June 2017

What Counts as Desiring the Actual Good?

Sean Clancy
Syracuse University

Abstract

Here is a simple observation about moral character: Moral virtue apparently consists, at least in large part, in *caring about* the right things. When we imagine a virtuous agent, we find that she cares about particular considerations, and that her caring is at least part of what makes her virtuous. One cannot be fully virtuous, for example, unless one cares at least somewhat about the welfare of others. Here is a corollary: At least sometimes, agents are morally vicious because they do *not* care about the right things. An agent who just *doesn't care* whether others live or die should, for example, strike us as severely vicious.

And here, from Hume, is a simple observation about moral responsibility: In order for *an agent* to be blameworthy or praiseworthy for an action, that action must reflect something *about* that agent. This observation, too, is supported by common intuitions. Agents seem to be blameworthy *when and because* their actions reflect something *bad* about their moral character, and they seem praiseworthy *when and because* their actions reflect something good about their moral character. And we are generally reluctant to attribute blameworthiness in cases in which circumstances prevent an agent's character from being reflected in his actions – we typically excuse agents whose bad actions result from delusions or uncontrollable impulses, for example.

Here, finally, is an appealing synthesis of these observations. Agents are blameworthy for actions that reflect their moral vices, and moral vices consist, at least in large part, in having the wrong attitudes towards certain considerations. Therefore, it seems that agents are blameworthy when their actions reflect such attitudes. And, since virtues consist, at least in large part, in having the *right* attitudes towards certain considerations, agents will be praiseworthy for actions that reflect *these* attitudes. This synthesis is also intuitively plausible. An agent who stands idly by and watches a child drown seems not only *vicious* in virtue of his indifference to human life,

but *blameworthy* in virtue of the fact that this indifference is reflected in his action. And an agent who makes significant sacrifices to help others is not only *virtuous* in virtue of her great concern for others, but also *praiseworthy* when she exercises her virtue.

The preceding observations raise two obvious questions: *Which* considerations are relevant to virtue and moral worth, and *which* attitudes are the "appropriate" ones to have towards these considerations? A recently-influential family of views (Arpaly 2002, 2003, 2006; Markovits 2010, 2012, Arpaly and Schroeder 2014a) offers a procedure for answering these questions. The considerations relevant to virtue and moral worth, according to these views, are the considerations that the correct normative theory identifies as relevant to determining the deontic status of an action, and the appropriate attitude towards a particular consideration is determined by that consideration's role as right-making or wrong-making. Thus, a virtuous agent will have positive or pro- attitudes towards those considerations that make actions good or right, and negative or anti- attitudes towards those considerations that make actions bad or wrong. Call accounts of this kind *actual good* (AG) accounts. A number of considerations count in favor of AG accounts. As noted, they do an excellent job of accommodating several intuitively plausible observations about character and moral worth. They are also equipped to provide intuitively plausible attributions of moral worth in a range of important cases.

But there are additional desiderata for an account of virtue and moral worth. Attributions of moral worth are not merely of theoretical interest but also of practical importance, as they are likely to have implications for which agents we should reward or punish. And while the correct attributions of virtue and moral worth seem to be obvious in some cases, they are *not* obvious in others. In particular, there are a number of socially, legally, and morally important cases of wrongdoing in which it is not intuitively clear how we should evaluate the agents in question.

These include the case of the psychopath; they also include cases of ideologically-motivated agents who act badly as the result of false moral beliefs. Preferably, our account of virtue and moral worth would be *useful* in guiding our judgments of moral worth in these difficult, real-world cases. Ideally, it would be *complete*, in the sense that it would offer a generalized procedure for assessing moral worth in *all* cases: Our account would take the correct normative theory as an input, along with the attitudes reflected in an agent's action, and then act as a function that outputs an unambiguous judgment of moral worth.

I argue that existing AG accounts are *not* complete in this sense, as there are realistic problem cases in which these accounts struggle to provide an unambiguous judgment of moral worth. That there are such cases at all means that there is a theoretical problem, and that we do not yet have a complete account of moral worth. That some of these cases are realistic means that there is also a practical problem, as these are precisely the cases in which we may need to rely on our account to guide our judgments. The reason that certain cases are problematic, briefly, is that normative theories identify a range of features of actions as right-making and wrong-making. Because an action can reflect appropriate attitudes towards some of these features while reflecting inappropriate attitudes towards others, our account will produce different attributions of moral worth depending on *which* of an agent's attitudes we evaluate him against.

Fortunately, I argue, this problem can be solved. It will require us to develop a further procedure for determining which attitudes, towards which right- and wrong-making features, we should use to evaluate agents. This in turn will require us to address a further substantive question as to which kinds of attitudes *count*, for the purposes of assessing character and moral worth, as appropriate or inappropriate attitudes towards that which is actually good or bad. Once this work has been done, however, we will have an account of moral worth that is much more

powerful, and that is able to provide unambiguous judgments in the cases which were previously problematic. This strengthened account has potentially surprising consequences when applied to the real world, implying, for instance, that psychopaths are morally blameworthy, and that many seemingly well-meaning agents are morally vicious.

What Counts as Desiring the Actual Good?

Sean Clancy

B.A. Princeton University, 2009
M.A. University of Maryland, 2011

A Dissertation
Submitted in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Philosophy

Syracuse University
May 2017

Acknowledgments

This dissertation would not have been possible without the support of my friends, family, and colleagues. I cannot hope to mention by name every person who has helped me in one capacity or another over the long duration of this project, but I will single out the following individuals for special thanks:

– My advisor Ben Bradley, whose support at every stage of this project was indispensable in bringing it to fruition.
– Hille Paakkunainen, who provided comments on countless papers which ultimately served as the groundwork for this dissertation.
– My other committee members: Nomy Arpaly, Mark Heller, and David Sobel.
– My many colleagues, at Syracuse and elsewhere, who have provided formal or informal feedback on various aspects of this project; in particular Dante Dauksz, Joe Hedger, James Lee, Isaiah Lin, Scott Looney, Amy Massoud, Dan Moller, Byron Simmons, Walter Sinnott-Armstrong, Steve Steward, Matt Talbert, Travis Timmerman, and Preston Werner.
– My parents, Kevin and Kathleen.

Table of Contents

Chapter One

Moral Worth and the Actual Good

I. Introduction

The *moral worth* of an action is a measure of the moral blame or credit that an agent merits by performing it. Agents are *praiseworthy* for performing actions with positive moral worth, and *blameworthy* for performing actions with negative moral worth. The moral worth of an action is distinct from its *deontic status* as right or wrong; there is room in conceptual space for agents who are praiseworthy for performing wrong actions, or blameworthy for performing right actions. Similarly, the moral worth of an action is apparently independent from what the agent *believes* about that action's deontic status. Some agents, such as those who act compassionately against their better judgment, seem intuitively to be praiseworthy even though they believe their actions to be wrong. Others, such as ideologically-motivated war criminals, seem to be blameworthy even though they believe their actions to be right.

In recent years, Nomy Arpaly (2002, 2003, 2006), Julia Markovits (2010, 2012), and Nomy Arpaly and Timothy Schroeder (2014a) have offered accounts of moral worth that are particularly well-equipped to accommodate our intuitions about such cases. These accounts differ in their details, but all are based on the plausible observation that agents generally seem to be praiseworthy when and because they desire – or respond to, or are motivated to pursue – those things that are actually morally good. Conversely, agents generally seem to be blameworthy when and because they *fail* to desire – or respond to, or be motivated to pursue – that which is actually good. Call accounts of this kind *actual good* (AG) accounts of moral worth.

One advantage of AG accounts is their aforementioned ability to provide intuitively

plausible attributions of praiseworthiness and blameworthiness in certain cases. They explain, for instance, why Huckleberry Finn is praiseworthy for helping his friend to escape from slavery, even though he falsely believes that it is wrong for him to do so. Huck is concerned about his friend's well-being, which is (presumably) *actually* morally important; thus, his actions reflect a pro-attitude towards the actual good.[1] AG accounts can also explain why ideologically-motivated war criminals are blameworthy even though they may believe themselves to be acting rightly. Because death and suffering are *actually* morally bad, war criminals show their lack of aversion to the actual bad when they act.

I believe that the ability of AG accounts to handle these cases in an elegant and intuitively plausible way counts strongly in their favor, and that some AG account is likely to be correct. The discussion in this dissertation is motivated by the fact that there are *other* cases which existing AG accounts do *not* handle well. In a range of interesting cases, existing AG accounts struggle to produce unambiguous attributions of moral worth. This is a problem, especially given that some of these cases are realistic. Moral worth is often of practical importance, affecting, for instance, which agents it is appropriate to punish. Ideally, an account of moral worth would be useful for guiding our judgments about real-world cases in which it is not intuitively clear whether an agent is blameworthy for his actions. One goal of this dissertation, therefore, is to propose a solution that will allow AG accounts to produce unambiguous judgments of moral worth in the full range of interesting and realistic cases.

However, this problem is not merely of practical importance, nor is the solution merely a matter of tweaking existing accounts so as to accommodate additional cases. For, I argue, the

1 For an extensive discussion of the moral worth of Huckleberry Finn's actions, see Arpaly (2003) pp.75-8; also Arpaly and Schroeder (1999) and (2014a) pp.178-9.

ambiguous judgments that AG accounts produce in some cases are symptomatic of a more fundamental problem that has not previously been appreciated. The ambiguity results from an assumption regarding the relationship between normative theories, on the one hand, and the features of actions that make them right or wrong, on the other. It generally seems to have been assumed that a normative theory identifies, in a fairly straightforward way, a very limited number of features that are right-making and wrong-making. I argue that this assumption is false, and that the limitations of previous AG accounts are due to their implicitly incorporating this faulty assumption. A second major goal of this project, therefore, is to uncover an important problem for those interested in the relationship between normative theories, right- and wrong-making features, and normative explanation.

Once this false assumption is discovered and rejected, and a bit more work is done, AG accounts will have a much broader scope of applicability, with the ability to provide unambiguous attributions of moral worth in a much wider range of cases. The third major goal of this dissertation is to explore the sometimes-surprising implications of these newly-strengthened accounts. They imply, for instance, that psychopaths are blameworthy for their bad actions. They imply that agents are often blameworthy when they take the wrong position on controversial moral questions, such as the moral status of animals or the moral permissibility of abortion. Finally – with the aid of a minor extension – they also imply that agents are blameworthy when their actions reflect moral concern that is directed at inappropriate targets.

This dissertation is divided into two parts; the first is concerned primarily with describing the difficulty for AG accounts, diagnosing its dependence on a problematic assumption about normative theories, and proposing a solution. In the remainder of this chapter, I describe some initial assumptions and general motivations for this project; offer some reasons for thinking that

an AG account of moral worth is likely to be correct; and dispense with some preliminary objections. In Chapter Two I describe the problem cases that existing AG accounts are unable to handle properly, and in Chapter Three I offer a diagnosis for this problem: their failure results from a faulty assumption about the relationship between normative theories and the right- and wrong-making features of actions. In Chapter Four I propose and defend a solution, which requires us to to identify a particular subset of right-and wrong-making features as the ones *relevant* to assessing moral worth.

The second part of this project concerns the implications of the solution defended in Chapter Four; with the problematic ambiguity eliminated, our newly-strengthened AG account has sufficient power to provide unambiguous judgments of moral worth in a range of previously problematic cases. In Chapter Five, I first illustrate how my strengthened account can resolve the problem cases discussed in Chapter Two, and then provide a brief overview of its other implications – in short, it implies that many agents may be morally much worse, and much more blameworthy for their actions, than we initially believed them to be. In Chapter Six, I discuss the case of psychopaths, who our new account tells us are blameworthy. Since many have argued that psychopaths are *not* blameworthy, this implication will need to be defended; I therefore offer an extended defense against one of the most important arguments that psychopaths are excused from blame. In Chapter Seven, I turn to a family of puzzling cases which have not previously attracted significant attention and which involve agents who act badly as the result of moral concern which is directed at inappropriate targets. I argue for a modest extension to the AG account defended in earlier chapters, and show that this extension implies that the agents in these cases – who respond to irrelevant considerations *as though* they provide moral reasons – are also vicious and blameworthy.

II. What is Blameworthiness?

Praiseworthiness and blameworthiness are *not*, I take it, the concepts *most* fundamental to our moral thinking. The most fundamental concepts are presumably either the *right* and the *wrong* – concerning which actions we have reason to perform – or the *good* and the *bad* – concerning which actions or states of affairs have positive or negative value. I make no assumptions here as to which, if either, of the right or the good is *more* fundamental. In fact, for simplicity I will later use these terms more or less interchangeably, generally subsuming both the good and right under the term "good," and both the wrong and the bad under the term "bad." Much of moral philosophy is premised on the assumption that some actions really *are* right or wrong, and that some states of affairs really *are* good or bad; in other words, on the assumption of moral realism. I share this assumption here. I further assume that there is a single, correct normative theory, although I make no substantive assumptions about the *content* of this theory. Nor do I make any substantive assumptions about the nature of moral properties, e.g. as to whether they are identical to natural properties or not. Finally, I assume that whatever it is that is ultimately morally good or right, its goodness or rightness is *not* fundamentally a *relational property*; thus I reject relativist or subjectivist views of morality.

Given their foundational role in moral thinking, judgments about right and wrong (or good and bad) are likely to elicit the strongest moral *intuitions*. Our confidence about certain claims regarding the rightness or wrongness (or the goodness or badness) of certain actions or states of affairs should be as high as our confidence in any other claims about moral philosophy. Consider:

(1) The Nazis acted wrongly when they carried out systematic campaigns of extermination.

(2) *Ceteris paribus*, a world in which large groups of innocent people are systematically exterminated is morally worse than one in which they are not.

I take it that most of us will have high pretheoretical confidence in (1) and (2), and in many other claims about the wrongness of particular actions or the badness of particular states of affairs. In fact, I suspect that the case for moral realism in general is motivated in large part by the strength of our intuitions regarding claims like (1) and (2) – many of us are so confident that certain things *are* wrong or bad that we conclude on this basis that there must *be* such things as wrongness or badness. So I take it that rightness and goodness (or wrongness and badness) claims about "obvious" moral facts – claims like (1) and (2) – form the "bedrock" of our intuitive picture of morality. To abandon these claims would be severely revisionary, and we have reason to be reluctant to do so.

I propose here that certain claims about blameworthiness and praiseworthiness are *nearly* as foundational to moral thinking. Consider:

(3) Given certain modest assumptions about the psychology of the Nazi leadership (e.g. that they were not being controlled, like puppets, by extraterrestrials), they were morally blameworthy for organizing and carrying out systematic campaigns of extermination.

My pretheoretical confidence in (3) is almost as strong as my pretheoretical confidence in (1) and (2). I suspect that many others will feel similarly. This, I think, gives us a fairly strong *prima facie* reason to be realists about moral worth. There might be good theoretical reasons to reject (3), such as, for instance, if no one has the free will necessary for moral responsibility. But, like the rejection of moral realism itself, this would be a severely revisionary position, and we should be reluctant to accept it.

Other kinds of claims may also be located close to the foundation of our moral thinking.

For example:

> (4) Again assuming the truth of certain modest assumptions about their psychology, the Nazi leadership deserved to be punished for organizing and carrying out systematic campaigns of extermination.

As with other claims, we might have theoretical reasons to reject this one; it might, for instance, turn out that no one has free will, or that desert fails to exist for some other reason. But it certainly seems that, if *anyone* deserves to be punished, it must be the Nazi leadership. Pretheoretically, I think, we should be nearly as confident in (4) as we are in (1) through (3). And thus we should be reluctant to abandon it unless compelled to do so.

I believe that we are *not* compelled to do so – that agents sometimes *are* praiseworthy or blameworthy for their actions, and that they *do* sometimes deserve to be punished or rewarded for the actions they perform. A full defense of these claims is beyond the scope of this project, so I state them merely as assumptions: I assume that agents are sometimes blameworthy or praiseworthy for their actions, and that agents sometimes deserve to be punished or rewarded on the grounds of having acted in ways for which they are blameworthy or praiseworthy. The first assumption will *need* to be accepted, as least as a stipulation, in order for the discussion in the remainder of this dissertation to make sense – if there are no such things as praiseworthiness and blameworthiness, this project will not get very far.

The second assumption – that agents sometimes deserve rewards or punishments on the basis of their being blameworthy or praiseworthy for their actions – is less essential to this project. Questions concerning the proper extension of praiseworthiness and blameworthiness may be of theoretical interest even if they have no implications for how agents deserve to be treated. This assumption does, however, serve two purposes. First, it helps to motivate the problem described in the next chapter. I will argue that existing AG accounts fail to make

unambiguous attributions of moral worth in certain cases, some of which are realistic. If agents deserve to be punished or rewarded as the result of their praiseworthiness or blameworthiness, and if we hope to be able to rely at least in part on our best theoretical account of moral worth for guidance as to which agents to reward or punish, then this problem will be of *practical* significance.

Second, this assumption is useful to illustrate the *kind* of moral blameworthiness and praiseworthiness that I have in mind in this dissertation. "Moral responsibility" is sometimes disambiguated in various ways; David Shoemaker (2011, 2015), for instance, distinguishes between responsibility as *attributability*, responsibility as *answerability*, and responsibility as *accountability*. These are ordered from "weakest" to "strongest", with accountability representing the kind of responsibility required for agents to be held *to account* for their actions. When I assert that an agent is *blameworthy* for an action, I mean to assert that he is responsible in the *strongest possible* sense; that is, in the sense that could potentially ground his deserving punishment.[2] Whether agents *actually* deserve punishment is not central to this dissertation, and the primary discussion could be conducted even under the assumption that, for whatever reason, no one deserves to be punished. But it is essential to be clear that I am concerned with a "full-strength" conception of blameworthiness here – the *kind* of blameworthiness that *would* ground desert, *if* desert existed.

Because this project is partly motivated by the need to determine whether punishment is appropriate in certain difficult cases, blameworthiness will play a more prominent role than praiseworthiness in the following discussion. The AG accounts discussed shortly, however, are

2 Note that in Shoemaker's (2015), it is less clear that accountability is intended to be a "stronger" kind of responsibility than the others, as Shoemaker argues that some agents can be accountable without being answerable. In any case, the kind of responsibility of interest here is the kind that could in principle ground desert, regardless of whether this kind is properly understood as the "strongest" kind.

meant to be unified accounts of both blameworthiness *and* praiseworthiness, offering a general procedure that allows us to assess agents for either. For brevity, I will often simply refer to blameworthiness; unless otherwise specified, however, it should be noted that I intend claims about blameworthiness to be applicable, *mutatis mutandis*, to praiseworthiness as well.

So far as further assumptions about the *nature* of blameworthiness, as well as the act of blaming, I wish to remain as neutral as possible for the time being.[3] Two clarifications, however, are important. First, I take it that, strictly speaking, agents are blameworthy *for* actions. For there to be blameworthiness, there must be an agent, and that agent must have performed an action. We can still write intelligibly of blameworthy *agents* and blameworthy *actions* in isolation – and I will do so at times in this dissertation – but this should not be taken to imply that one can exist without the other. When I write that an *agent* is a blameworthy agent, I mean that he has performed some action for which he is blameworthy; when I write that an action is a blameworthy action, I mean to imply that some agent is blameworthy for it.

Second, I am assuming that blameworthiness is *not* contingent upon any human blaming *practices* nor upon any human *dispositions* to blame in certain ways. Strawson (1962) famously ties the aptness of blaming an agent to that agent's being an apt target for certain "reactive attitudes", such as resentment. On what I take to be the dominant interpretation of this paper[4], Strawson is asserting that our actual blaming practices – or our dispositions to engage in certain blaming practices – *determine* which agents are blameworthy. That is, human psychology is such that we are disposed to blame agents with certain properties; it is in virtue of both our psychology and the presence of these properties that these agents are blameworthy. If our

3 I will revisit these questions in Chapter Four.
4 See, e.g. Eshleman (2014).

psychology were different, such that we were disposed to blame agents with different properties, then the conditions for blameworthiness would also be different. But Strawson's view amounts to what I consider to be an unacceptably deflationary account of blameworthiness; it seems to be analogous to relativism with respect to moral truths, and suffers from many of the same problems. It does not accommodate the possibility that most humans could be mistaken about which conditions make an agent blameworthy. It rules out the possibility of surprising new discoveries about which agents are blameworthy – some of which I will argue for in subsequent chapters. And it does not give blameworthiness as much metaphysical "heft" as seems appropriate, reducing it from a genuine, intrinsic property of agents to an extrinsic, relational one.

So, to recap: I assume that there are some objective facts about which actions are right and wrong, and/or which states of affairs are good or bad. I assume that some agents are blameworthy or praiseworthy for performing certain actions. I assume that blameworthiness and praiseworthiness are not contingent on social practices or human psychology. And I use "blameworthiness" to designate the "strongest" form of blameworthiness, the kind that could in principle ground an agent's deserving punishment. I have not defended these assumptions at length; my intention here is merely to set the stage for the discussion of moral worth that follows. In the remainder of this chapter, I turn from questions of the nature of praiseworthiness and blameworthiness to the question of their extension. Which agents are praiseworthy and blameworthy for which actions, and why?

III. Actual Good Accounts of Moral Worth

I contend that this question is best answered by *actual good* (AG) accounts of moral worth. First, to avoid confusion, a note about terminology: Accounts of the kind discussed here

have sometimes been referred to as "attributionist" accounts.[5] As we will see, one condition for responsibility on these accounts is that an action reflect an agent's attitudes and thus be attributable to him. However, I choose to use a different term here – "AG accounts" – for three reasons. First, I wish to avoid confusion between the accounts of interest and the view of responsibility *as* attributability; as noted, I am interested in "full-strength" responsibility, and some, like Shoemaker, have treated attributability as a weaker kind.

Second, the feature of these accounts that is most salient to the following discussion is not their focus on whether an action is attributable to agent, but rather their account of *which* attributable attitudes make agents blameworthy or praiseworthy. The feature of interest here is that these accounts evaluate actions based on whether they reflect appropriate or inappropriate attitudes towards the actual good or actual bad. The central controversy discussed in this dissertation concerns the question of what counts as an attitude towards the actual good or bad, and thus it seems more appropriate for my purposes to emphasize this aspect of these accounts.

Finally, it seems that attributionist and AG accounts are not identical in their extension; Some attributionist accounts do not afford a central role to the actual good. T.M. Scanlon's (2008), for example, evaluates agents on the basis of whether their actions reflect an attitude that damages their potential for relations with others; it does not require us to evaluate agents based on their attitudes towards that which is identified as good by the correct normative theory.[6] And it might also be possible to construct an account that appeals to the actual good without being

5 See, e.g. Levy (2007), who characterizes Arpaly's (2003, 2006) as such. Talbert (2008) notes that the term "attributionist" is more frequently used by critics of these accounts than by supporters. Since my goal is to make these accounts stronger rather than to undermine them, this provides another reason to use a different term.

6 In Chapter Four, I argue that Scanlonian and AG accounts are compatible with one another, on the condition that it is inappropriate attitudes towards the actual good and bad that damage our potential for relations with others. For now, it is sufficient to note that a Scanlonian account *need not* be an AG account, and thus that not all attributionist accounts appeal to the actual good.

attributionist – although I am unaware of any such accounts having been discussed in the literature. In any case, it seems best to introduce a new term here, since only a subset of "attributionist" accounts are of interest.

I begin by describing a representative AG account in some detail. Existing AG accounts are similar in the respects that are important to this dissertation – they are all afflicted by the fundamental problem described in the next chapter, for instance – but space constraints prevent me from discussing all such accounts at length. I focus on Arpaly and Schroeder's (2014a) account here, because I believe that it most clearly illustrates the workings and motivations of accounts of this kind.

Arpaly and Schroeder offer the following:

> Praiseworthiness: a person is praiseworthy for a right action A to the extent that A manifests an intrinsic desire (or desires) for the complete or partial right or good (correctly conceptualized) or an absence of intrinsic desires for the complete or partial wrong or bad (correctly conceptualized) through being rationalized by it (or them).
>
> Blameworthiness: a person is blameworthy for a wrong action A to the extent that A manifests an intrinsic desire (or desires) for the complete or partial wrong or bad (correctly conceptualized) or an absence of intrinsic desires for the complete or partial right or good (correctly conceptualized) through being rationalized by it (or them).[7]

As we will see in subsequent chapters, the question of what counts as a desire for the actual good or bad is a complicated one. But a simple example will illustrate the general intention of this account. Suppose that the correct normative theory identifies one thing as morally bad – pain – and one thing as morally good – the absence of pain. A sadistic agent, who desires that others feel pain, would desire the bad. If he acted on this desire by actually inflicting pain, then his action would reflect this desire, and he would be blameworthy. Conversely, an

7 2014a, p.170.

agent who strongly desired that others *not* be in pain would desire the good; if this desire moved her to prevent others from feeling pain, she would be praiseworthy.

Two technical details of the account can be glossed over quickly. First, the requirement that a desire for the good or bad be *intrinsic* is intended to exclude cases in which agents desire something that is good or bad as a means to some other end.[8] Pain may be morally bad, but I am clearly not blameworthy if I cause my friend pain in order to wake him from a coma. And saving lives may be morally good, but I am not praiseworthy if I only want to save lives as a means to becoming famous. In the cases discussed in this dissertation, unless otherwise specified, I mean to assume that the relevant desires possessed by the agent in question are intrinsic ones. Second, the requirement that an action *manifest* a desire that *rationalizes* it can be understood, for our purposes, as a requirement that the action be caused by the desire in the right sort of way.[9] The details of manifestation and rationalization do not affect the problems that I describe later, and I will assume, in each of the cases discussed, that the agent's action does appropriately manifest the relevant desires. To indicate that I am abstracting away from the specifics of Arpaly and Schroeder's account of manifestation, I will continue to refer to an action as *reflecting* a desire or attitude of a certain kind.[10]

Two other features of this account require additional commentary, as they are particularly important to the following discussion of problem cases. First, Arpaly and Schroeder distinguish between complete and partial goods, and claim that an agent is praiseworthy for an action that reflects a desire for a good of either kind. The *complete* good consists in the entirety of what the correct normative theory prescribes. If hedonistic utilitarianism is correct, for example, the

8 Ibid., p.6.
9 Ibid., pp.170-1.
10 The details of how we understand the reflection requirement may determine whether or not an AG account is implicitly compatibilist; I return to this subject in the final section of the present chapter.

complete good consists in maximizing the balance of pleasure over pain. A *partial good* is something that we have a *pro tanto* moral reason to pursue, given the truth of the correct normative theory.[11] Supposing again that hedonistic utilitarianism is correct, one might desire the partial good by desiring that a particular person be saved from pain. The distinction between complete and partial goods is particularly relevant to the following discussion, because it allows Arpaly and Schroeder's account to make plausible and unambiguous attributions of moral worth in *some* cases which would otherwise be problematic. The distinction makes it possible for their account to judge an action praiseworthy if it reflects a desire that a particular person be saved from pain, even if it does not reflect a desire that overall utility be maximized. This is an intuitively plausible attribution of praiseworthiness, and it is indeed desirable that an account of moral worth should provide it. In the next chapter, however, I argue that the problem facing AG accounts is more complex than this, and that the distinction between complete and partial goods does not provide an adequate solution.

Second, Arpaly and Schroeder specify that a desire for the good or bad must be correctly *conceptualized* in order for it to affect moral worth. The idea is apparently that the correct normative theory not only identifies certain things as good and bad, but identifies them using certain concepts; and, in order for a desire to affect moral worth, it must be a desire for the good or bad under the same concepts as those employed by the normative theory.[12] Hedonistic utilitarianism, for example, identifies pleasure as good and pain as bad; significantly, however, it identifies the good as pleasure *under its description as pleasure* and the bad as pain *under its description as pain*. It is possible for an agent to desire these things under different descriptions.

11 Ibid., pp.165-6.
12 Ibid., p.15, pp.176-8.

If, for instance, it turns out that pleasure is identical to certain neural events, then an agent could desire the good, in some sense, by desiring that certain neural events be maximized. But this would not be a desire for the good under the correct conceptualization, as the correct normative theory commands us to promote pleasure under its description as pleasure rather than pleasure under its description as certain neural events. This conceptualization requirement is apparently intended, at least in part, to exclude these sorts of desires from affecting moral worth.[13] And this, too, is a desideratum of an account of moral worth, for it seems clear that an agent would not be praiseworthy or blameworthy for promoting certain neural events without realizing that they instantiated pain or pleasure. Conceptualization is particularly important to the following discussion, because it may seem that the problem cases I present can be resolved by asking whether the agents in question desire the good under the correct conceptualization. However, as with the distinction between complete and partial goods, I argue that the appeal to the correct conceptualization of the good cannot adequately resolve the full range of problem cases.

IV. Why an Actual Good Account is Likely to be Correct

In the chapters that follow, I will be discussing a previously-unappreciated problem for existing AG accounts. My ultimate aim, however, is not to argue against accounts of this kind, but rather to improve them by diagnosing and correcting this problem. I think that some AG account is likely to be correct, and hope that the strengthened account that I defend in later chapters will bring us closer to the true account of moral worth. Before moving on, it will be useful to discuss several of the considerations in favor of AG accounts. This will help to motivate the following discussion by showing why we should take the time to solve the problem facing

13 Ibid., 166-7. This conceptualization requirement is also intended to exclude desires for the good *de dicto* – that is, desires for the good under its description *as* the good. I discuss the relevance of the good *de dicto* shortly.

these accounts, rather than simply abandoning them altogether.

AG accounts are broadly Humean, in that they assess actions on the basis of what they reflect about an agent's moral character. In order for it to be appropriate for us to blame or praise *an agent* for an action, Hume famously observes, the action must be due to something that is *wrong with* or *right with* that agent – a defect or excellence in his character.[14] AG accounts further specify that the quality of an agent's moral character depends on the quality of his *attitudes* – or, more specifically in the case of Arpaly and Schroeder's account, on his desires – and this should seem intuitively plausible. It seems quite natural to think that wanting certain things can make agents morally bad or vicious. Sadism is a paradigmatic moral vice, and it consists in the desire to cause pain to others. Failing to want certain things can also constitute a character defect: imagine an agent who care so little about others that he watches a child drown rather than ruin his shoes by rescuing her. Conversely, agents who want certain things very strongly seem to be virtuous, or morally excellent, as a result – we admire those agents who care so much about others that they are willing to make significant sacrifices to help them.

So AG accounts proceed from two very plausible claims: The moral worth of an action depends on what it reflects about the moral character of the agent performing it, and the quality of an agent's moral character depends on the contents of his desires. Perhaps the most surprising aspect of AG accounts concerns *which* desires they hold to be reflective of moral character and therefore relevant to moral worth: On these accounts, desires for the *actual* good and bad matter for moral worth, but desires for that which is *believed* to be good or bad do not.

An agent who desires to do that which is morally good, *whatever it may turn out to be*, is

14 *Treatise of Human Nature*, Book II, Part III, Section II.

sometimes described as desiring the good *de dicto*.[15] Archetypal AG accounts, such as Arpaly and Schroeder's, hold that desiring or failing to desire the good *de dicto* makes no difference either to an agent's moral character or to the moral worth of his actions – agents merit no praise for acting out of a desire for the good *de dicto*, nor do they deserve blame because their actions reflect a lack of desire for the good *de dicto*. This feature of AG accounts may initially be surprising, as we may have some intuitive inclination to praise agents for doing what they *believe* to be right, as well as to blame agents for doing what they *believe* to be wrong.

It is first important to note that this feature of AG accounts is intended to follow from the requirement that the good and bad be desired under the correct conceptualization, where the correct conceptualization is the one employed by the correct normative theory. To desire the good *de dicto* is to desire the good under its description *as* the good. But it seems unlikely, and perhaps impossible, that the correct normative theory simply commands us to promote the good under its description *as* the good – such a normative theory would be completely vacuous. So, whatever the correct conceptualization of the good for the purposes of attributing moral worth, it apparently cannot be a conceptualization of the good as good, and an agent's attitudes towards the good *de dicto* are irrelevant to moral worth.

Upon further reflection, I think, many of us will find that this feature is more intuitively plausible than we may initially have believed. It is, for instance, what enables AG accounts to correctly handle cases such as those described in the introduction. Huck Finn cares deeply about the welfare of his friend, but does *not* care as much about the good *de dicto* – after all, he decides to help his friend even though he believes that to do so would be wrong. But Huck's lack of concern for the good *de dicto* does not seem to diminish his moral character, nor to make him

15 See, e.g. Smith (1994) pp.71-6, 82-3.

blameworthy. And, as Arpaly and Arpaly and Schroeder point out, *if* Huck were to decide differently – because his desire for the good *de dicto* were stronger than his desire to help his friend – this would seem to indicate a defect of moral character rather than an excellence.[16]

This is further illustrated by a more dramatic case – a war criminal might believe that he has a moral obligation to promote ethnic homogeneity, and this might cause him to ignore the suffering that he inflicts on his victims. The war criminal's actions reflect a desire for the good *de dicto* – he is trying to do that which he *believes* to be morally right – but this desire does not seem to be a credit to his moral character, nor does it seem to excuse him from blameworthiness. When we consider these kinds of cases, it should seem clear that moral character depends, at least in large part, on one's attitudes towards such things as the suffering of other beings – *not* on one's attitudes towards abstracta such as the right and the good.[17]

That moral worth depends on one's attitudes towards the good *de re* – rather than the good *de dicto* – has a corollary that will be relevant to the following discussion. On AG accounts, an agent's *non-moral* ignorance can excuse him from blameworthiness, while his *moral* ignorance cannot. By *moral ignorance*, I mean ignorance of the basic moral facts, such as facts about what is intrinsically morally good or bad, or about what considerations do or do not provide intrinsic moral reasons for action. So, supposing that lying is intrinsically morally wrong, an agent who is unaware *that* lying is wrong suffers from moral ignorance. Since I include false beliefs as instances of ignorance, an agent who positively believes that lying is *not*

16 See again Arpaly (2003) pp.75-8, Arpaly and Schroeder (1999), and Arpaly and Schroeder (2014a) pp.178-9.

17 AG accounts in their current form do a good job of handling *some* cases of ideologically-motivated wrongdoers – those cases in which these agents' actions reflect a lack of concern for that which is actually morally important. In the final chapter, I will argue that some cases of ideologically-motivated wrongdoing are not like this, and that such agents need not lack concern for that which is actually good, nor desire that which is actually bad. However, I argue that these cases can be accommodated by a further extension of existing AG accounts.

wrong also suffers from moral ignorance.

By *non-moral* ignorance, I mean ignorance of, or false beliefs about, any matters *other than* the basic moral facts. For our purposes, the interesting kinds of non-moral ignorance will concern ignorance of or false beliefs about either the features of one's actions or the background facts about the world that bear on the deontic status of actions. An agent might abhor harming animals, but falsely believe that cetaceans are not harmed by being kept in captivity. This agent's false belief about what does and does not harm cetaceans is an example of *non-moral* ignorance about the world. An agent might abhor causing pain, but falsely believe that the action he is performing does *not* cause pain; this, too, counts as non-moral ignorance.

The reason that non-moral ignorance can excuse, on AG accounts, is that it can cause agents to act badly without thereby displaying a desire for the actual bad or indifference towards the actual good. Suppose that an agent abhors harming animals, and gives a donation to the local aquarium's cetacean program under the belief that it will benefit these animals; as it turns out, his action enables the aquarium to continue inflicting harm on cetaceans. This agent's action does *not* reflect a desire to harm cetaceans, nor an indifference to harming them – the agent does not believe that he is harming cetaceans by acting, nor is he motivated to do so. He is trying to help cetaceans rather than harm them, and this is the desire that is reflected in his action; his ignorance, in this case, can be viewed as a kind of external obstacle that prevents him from successfully achieving his desires. Since non-moral ignorance can prevent an action from reflecting an agent's desires, it can excuse agents for actions that are bad. The excusing power of non-moral ignorance is analogous to that of other obstacles that might prevent an agent's actions from reflecting his attitudes. If, for instance, my physical disability prevents me from saving the drowning child, it excuses me from blame for my failure to do so. My failure to save the child

reflects only my disability, rather than my attitudes towards the child, and therefore does not reflect on my character and does not have (negative) moral worth.[18]

The same is not true of moral ignorance. Ignorance about what is good can, of course, prevent an agent from successfully acting on his desire to pursue the good *de dicto*. But this is irrelevant to moral worth because desires for the good *de dicto* are themselves irrelevant. It is true that the sincere Nazi, who wants to act well and who believes that ethnic cleansing is morally good, is led astray by his false belief. But the false belief does *not* act as a barrier between the Nazi's *character* and his actions, because one's character does not consist in one's attitudes towards the good *de dicto*. On AG accounts, an agent's moral character consists in his attitudes towards the actual good and bad. Part of the actual good, presumably, consists in not murdering people. Since the Nazi is aware that he is murdering people, and he fails to be deterred by this knowledge, his action *does* reflect a lack of concern for that which is actually morally important; thus it *does* reflect poor moral character, and he *is* blameworthy.[19]

The conclusion that moral ignorance cannot excuse may initially be surprising. But it is a corollary of the same feature that allows AG accounts to handle Huck-Finn- and war-criminal-type cases properly, and reflecting on such cases will, I expect, diminish the intuitive implausibility of this feature of AG accounts. An additional advantage of this feature is that it allows us to defeat an argument that would otherwise force us to adopt a skeptical position about blameworthiness. Gideon Rosen (2004) argues that we ought never to be confident in our

18 Note that while non-moral ignorance *can* excuse, nothing here is meant to imply that it must *always* excuse. Agents who came to have their false beliefs (or to lack true beliefs) through epistemic irresponsibility or self-deception may well be blameworthy for their resulting bad actions.

19 Real-world cases are complex, and many real-world wrongdoers possess a constellation of false moral *and* non-moral beliefs which may be closely associated with one another. In Chapters Five and Seven, I turn to the analysis of complex, realistic cases in more detail. I will ultimately argue that AG accounts *should* attribute blameworthiness to real-world Nazis and many other ideologically-motivated bad actors. But for now, the characterization of the sincere Nazi can be treated as an abstraction for the purposes of illustration.

attributions of blameworthiness to others. He begins with the assumption that agents cannot be blameworthy for doing what they believe to be right; he makes the further, empirical claim that agents very often do believe themselves to be acting rightly, and that instances of genuine *akrasia* are very rare. Since we cannot reliably tell when another agent is acting akratically, we ought to assume that other agents believe themselves to be acting well when they act badly, and we ought not to blame any of them.

An obvious response to this argument is to appeal to epistemic sins on the part of the agents who act badly – it might be claimed that while Nazis and other wrongdoers may believe themselves to be acting rightly, their implausible moral beliefs must be the result of epistemic mismanagement at some point in the past. We can then blame these agents either for their original act of epistemic mismanagement, or claim that they are derivatively blameworthy for the bad acts that result from this culpable ignorance. But Rosen blocks this strategy by applying the same argument to bad epistemic actions. An agent cannot be blameworthy for an epistemic action, even if it is an irresponsible one, if the agent *believes* the act to be epistemically responsible. And acts of epistemic *akrasia* are also likely to be extremely rare, as agents are unlikely to deliberately manage their beliefs in an irresponsible way. Ultimately, Rosen claims, an agent can only be blameworthy if there is an instance of "clear-eyed akrasia" somewhere in the causal history of his action. Because akrasia is rare, and because we are very rarely in a position to attribute it to other agents, we ought to refrain from blaming others.

Rosen's argument seems likely to succeed if we grant the assumptions that he requires. Furthermore, the empirical assumption – that clear-eyed akrasia is rare in human agents – seems accurate. Akrasia may exist, but it is not likely to be responsible for many of the most spectacular acts of evil – certainly some of the worst agents, and the ones that seem most appropriate for us

to blame, are motivated by their ideologies, and thus are pursuing what they take to be good. And akrasia seems likely to be just as rare when it comes to belief management – agents rarely *decide* to deceive themselves or to refrain from thinking things through. So Rosen's argument apparently poses a real danger of forcing us to abandon claims like (3) and (4) – it might force us to withhold judgment on some of the worst agents in history.

But Rosen's argument depends crucially on the assumption that agents cannot be blameworthy for actions that they believe to be right; if we reject this assumption, we cut the argument off at the beginning. As noted, this assumption will be false if an AG account is correct. On AG accounts, agents are blameworthy for actions that reflect the wrong attitudes towards the *actual* good and bad. And, as established in the previous discussion, many wrongdoers will display the wrong attitudes towards the actual good and bad even as they display a desire for the good *de dicto*; the Nazi, for example, displays contempt for human life and is thereby blameworthy. So while the implication that moral ignorance cannot exculpate may surprise us, I believe that it is ultimately a consideration in favor of AG accounts, as it is what enables these accounts to neutralize a powerful argument in favor of skepticism about blameworthiness.[20]

V. Free Will and the Unity of the Virtues

In the chapters that follow, I present what I take to be a serious and complex difficulty for AG accounts of moral worth. Before doing so, it will be helpful to address two preliminary worries about AG accounts that I believe can be resolved more easily. First, we might wonder whether AG accounts require us to assume a compatibilist view of free will. While such an

20 Elizabeth Harman points out the power of AG accounts in this regard, in her (2011) response to Rosen. I discuss Harman's paper at greater length in Chapter Seven.

assumption would not necessarily be problematic, it would of course limit the potential interest of these accounts to those who are at least willing to seriously entertain compatibilism. Second, while the account of moral virtue that underlies AG accounts is intuitively plausible in many respects, we might worry that it is too simplistic. What if there are other components of moral virtue that do not consist in one's attitudes towards the actual good? If so, would these additional components of virtue imply that moral worth does not depend *solely* on which attitudes an action reflects towards the actual good and bad?

AG Accounts and Compatibilism

AG accounts of moral worth would appear to be particularly consonant with *source compatibilist* accounts of free will, on which an agent performs an action freely just in case he is the *source* of that action in the right sort of way.[21] On AG accounts, agents are praiseworthy or blameworthy just in case an action reflects something good or bad about their moral character. The requirement that an action reflect an agent's character could be understood as a requirement that the agent be the source of that action; because actions could reflect an agent's character even in a deterministic universe, this might lead us to conclude that AG accounts of moral worth implicitly depend upon or incorporate a compatibilist account of free will.

Whether or not a particular AG account of moral worth implies the truth of compatibilism depends on the details of that account, most importantly the details of the requirement that an action *reflect* an attitude. I have thus far not commented extensively on this requirement; the following is one possible way of filling it in:

> (5) An action reflects the character of the agent who performs it just in case the action is non-deviantly caused by the attitudes that constitute that agent's character.

21 See, e.g. Fischer and Ravizza (1998), Sartorio (2011).

If (5) is part of our AG account, then it seems as though this account will imply the truth of compatibilism. Since non-deviant causation by an attitude with the right content is sufficient for blameworthiness, and blameworthiness for an action presumably requires free will, it seems that non-deviant causation by an attitude would also be sufficient for free action. Since non-deviant causation by an attitude is possible even if an agent's actions are fully causally determined, this account of responsibility apparently implies that free will is compatible with determinism. My own view is that compatibilism is likely to be correct, so this implication is not, from my perspective, problematic.

But AG accounts need not require the truth of compatibilism for agents to be responsible. We are free to interpret the "reflection" requirement differently, so that non-deviant causation by an attitude is *necessary* but not *sufficient* for an action to reflect an agent's character. The full set of conditions for reflection might, for instance, be as follows:

(6) An action *reflects* an agent's character just in case it is *both*:
  (a) Performed freely; and
  (b) Non-deviantly caused by the attitudes that constitute that agent's character.

If (6) is part of our AG account, then that account will *not* imply the truth of compatibilism. Whichever conception of free will we prefer can be inserted into requirement (a), and it will constrain which actions can merit blame or praise. If we are incompatibilists who believe that only undetermined actions are free, we may read (a) as being satisfied only by those agents whose actions are undetermined.[22]

22 The incompatibilist might regard causation by an attitude as equivalent to causation by an event; if she is committed to the proposition that actions which are caused by events cannot be free, she might conclude that conditions (a) and (b) are never jointly satisfied. This would amount to the conclusion that agents are never responsible, but would not require a rejection of AG accounts as a correct description of what *would* be necessary *if* any agents were to be responsible. Alternatively – and plausibly – the incompatibilist might interpret causation by an attitude as distinct from causation by an event, in which case conditions (a) and (b)

Although Arpaly and Schroeder write at greater length about what is required for an action to "manifest" an attitude (their terminology for reflection), their concern is primarily with what counts as non-deviant causation, rather than with whether non-deviant causation by an attitude is a necessary or sufficient condition for the manifestation of that attitude.[23] On my reading, their account is compatible with either (5) or (6). In any case, my concern is not primarily with exegesis but instead with constructing the best AG account possible, and I see no theoretical reason to prefer (5) over (6).

So, for the remainder of this dissertation, it should be understood that I do not take AG accounts to require the truth of compatibilism, and leave it open for the reader to insert any conception of free will that she prefers. Note that the interesting problems discussed later in this dissertation center neither on what is required for an action to be free, nor on what is required for an action to reflect an agent's character. Instead, as I will make clear in the next chapter, my interest is in *what kinds of attitudes* count as good or bad *when* they are reflected in an agent's free actions.

AG Accounts and the Unity of the Virtues

As noted, AG accounts of moral worth are closely associated with a particular conception of moral character. On this conception, moral virtue consists in having the right attitudes towards the actual good and bad, while moral vice consists in lacking the right attitudes, or in having the wrong attitudes, towards the actual good and bad. Arpaly and Schroeder, at least, apparently intend for this to be an *exhaustive* conception of moral character; that is, they apparently contend that the quality of an agent's moral character depends *solely* on what his attitudes are towards the

could be jointly satisfied.

23 pp.61-72. Note that Arpaly and Schroeder require that an attitude "rationalize" – or provide a subjective reason to perform – an action in order for it to be "manifested" by that action. I have treated this as part of the non-deviant causation requirement.

actual good and bad. Thus we might say that on their view, the virtues are *unified* in one important sense – all virtues consist in correct attitudes towards the goods and bads identified by the correct normative theory. There is another important sense in which the virtues are *not* unified on their view; if there are multiple, independent goods, then there will be an independent virtue corresponding to each one. But the virtues apparently *are* meant to be unified in the sense that they are all attitudes, and that they are all attitudes towards the same *kind* of things – actual goods. Character traits such as wit and prudence are excluded; so are other kinds of attitudes, such as those towards the good *de dicto*.

What if we accept the general motivations for AG accounts, but are reluctant to accept this strong claim regarding the unity of the virtues? I have assumed, for example, that Arpaly's analysis of the case of Huckleberry Finn strikes many of us as compelling. It really does seem as though Huck is both virtuous and praiseworthy, and it really does seem as though his virtue consists in, and his praiseworthiness is grounded by, his desiring the well-being of his friend. This reaction seems to provide support for the claim that at least some significant part of virtue consists in desiring the actual good, and that agents can be significantly virtuous and significantly praiseworthy by virtue of having desires of this kind. But it does not necessarily support the claim that *all* of virtue consists in such desires, nor that such desires are the sole grounds for praiseworthiness. For the following also seems to be a reasonable intuitive reaction: While Huck Finn is virtuous and praiseworthy, he would be *more* virtuous and *more* praiseworthy if he *also* believed himself to be acting rightly. An agent can be good simply in virtue of desiring the actual good, we might think. But an agent is *better* if he desires *both* the actual good and the good *de dicto*.[24]

24 Hurka (2014) seems to be pressing an objection along these lines. More precisely, he claims that Huck Finn

Arpaly and Schroeder (2014b) briefly discuss and reject this possibility. Their argument against it seems to be premised on the claim that desires for the actual good and desires for the good *de dicto* are simply too different in kind for both of them to count as virtues. Regardless of whether or not this argument is convincing, it seems to presuppose that we are already committed to the idea that the virtues are unified in some way and therefore should all consist in the same sorts of attitudes; this is, of course, the very claim that might be challenged by an objector.

I believe, however, that we can accept an AG account, and that the discussion in this dissertation can proceed essentially unharmed, even if we do *not* accept the strong claim that *all* of virtue consists in having the right desires towards the actual good and bad. It is sufficient for our purposes if we accept that a *very significant part* of virtue consists in having desires of this kind. And these claims *are* supported by common reactions to cases such as Huck Finn and the sincere Nazi. Our reaction is that Huck Finn is fairly virtuous and fairly praiseworthy overall; this supports the claim that a *significant* part of virtue consists simply in desiring the actual good, and that an agent can be *significantly* praiseworthy if his actions reflect this desire. Similarly, our reaction is that the sincere Nazi is very vicious and very blameworthy overall; this supports the claim that a *significant* part of vice consists in failing to desire the actual good, and that an agent can be *significantly* blameworthy if his actions reflect the lack of such desires. It seems that a lack of desire for the good *de dicto* can be at worst only weakly vicious, as it does not make Huck Finn vicious overall. And it seems that a desire for the good *de dicto* can be at best only weakly virtuous, as it does not significantly absolve the sincere Nazi.

---

would be *less* virtuous if he didn't care about the good *de dicto* at all – that is, if he simply did what he believed to be wrong without feeling conflicted about it.

The claims made in later chapters do not require that one's attitudes towards the actual good and bad be the sole constituents of character and the sole basis for moral worth, but only that they be the most significant. I will later defend various claims about the moral worth of various actions, arguing, for instance, that psychopaths are blameworthy for their actions on the grounds that they reflect a lack of desire for the actual good. This argument is not undermined by the possibility that a psychopath's blameworthiness is slightly reduced by the fact that his action does *not* reflect a lack of concern for the good *de dicto*. So, the AG accounts of interest in this dissertation do not require that the virtues be unified in order to function; and, if readers prefer, they are free to reject Arpaly and Schroeder's strong claim that virtue is *exhausted by* desires for the actual good.[25]

To recap: Given the strength of their theoretical motivations, their ability to produce intuitively plausible attributions of moral worth in a number of cases, and their usefulness in avoiding a skeptical conclusion about blameworthiness, it seems that we have good reason to think that some AG account is correct. These accounts are furthermore not beholden to any particular view of free will nor to any strong claims about the unity of the virtues. In the next chapter, however, I argue that existing AG accounts are not prepared to make attributions of moral worth in the full range of cases encountered in the real world. AG accounts are correct, I think, in their claim that moral worth depends on whether an action reflects a desire for the actual good. The problem is that there is a further question which needs to be answered: What counts as a desire for the actual good?

25 In principle, an AG account could even be made to accommodate the possibility that traits such as wit and prudence are part of virtue, so long as they are a very *small* part. (A witty Nazi is clearly not *significantly* more virtuous than a non-witty Nazi.) This is unlikely to be necessary; as Arpaly and Schroeder (2016) make clear, their account is meant to address only *moral* virtue, and so non-moral excellences of character are simply outside of the scope of interest.

Chapter Two

What Counts as Desiring the Actual Good?

I. Problem Cases

Real-world decisions involving punishment are often influenced by attributions of blameworthiness. The precise relationship between blameworthiness and the appropriateness of punishment is controversial, and it is not my intention to resolve this controversy here. The following (very weak) principle, however, seems to be intuitively quite plausible:

(7) If an agent is *not* morally blameworthy for performing action P, this is a (moral) consideration *against* punishing that agent for P.

Some – utilitarians, for example – may be committed to the view that the appropriateness of punishment depends *solely* on its consequences. They will therefore reject (7). I expect, however, that almost everyone else will accept (7). It is consistent with a wide variety of views of punishment; for instance, it allows for the possibility that a number of factors may be relevant to whether it is suitable to punish an agent for a particular action. When weighing these factors, it should seem to most of us that the fact that an agent is *not* blameworthy for the action weighs *against* punishing him. Of course, (7) is also compatible with stronger claims about the relationship between blameworthiness and punishment. It is compatible with the claim that it is morally permissible to punish *only those* agents who are blameworthy; it is compatible with the even stronger claim that it is morally obligatory to punish *all and only* those agents who are blameworthy. It is not my intention to comment on the correct view of punishment here; if, as I suspect, (7) seems correct to most of us, that will be sufficient for the following discussion.

There are a number of real-world cases in which it may be unclear whether an agent is blameworthy for his action. One desideratum of an account of moral worth is that it provide

judgments of moral worth in these cases, so as to assist in guiding our decisions regarding the appropriateness of punishment. Of course, the moral worth of a particular action performed by a particular agent will depend on some questions beyond the scope of this project. It will depend on which desires are reflected in the agent's action, which is an empirical question about the psychology of a particular individual. Because AG accounts evaluate an agent based on his attitudes towards the actual good, we will also need to know the correct normative theory – which determines which things are actually good – in order to determine the moral worth of an action.

So a complete AG account of moral worth would not allow us to judge the moral worth of particular actions without first investigating these other questions. But, *if we knew* the answers to these questions, a complete account would take them as inputs and then output an unambiguous judgment of moral worth. That is, for each set of attitudes and normative facts, a complete account of moral worth would serve as a function mapping them to an attribution of praiseworthiness, blameworthiness, or neither. I describe three problem cases in this chapter which show that existing AG accounts are *not* complete in this sense. Even when we stipulate which moral facts obtain, as well as which desires are reflected in an action, there are still cases in which existing AG accounts do not provide clear judgments as to whether an agent is praiseworthy, blameworthy, or neither.

It is first important to be clear about what the problematic ambiguity is *not*: It is not a result of the fact that certain normative theories admit of multiple, distinct considerations that bear on the moral status of an action. A normative theory might, for instance, command us both to promote happiness and to promote scientific knowledge. In this case, there are two distinct actual goods – happiness and knowledge. And we can of course imagine an agent who desires

one of these without desiring the other; an agent might desire to find a cure for cancer only because it promotes happiness, and not because the cure would represent a new piece of scientific knowledge. Cases of this kind are not deeply problematic. AG accounts will presumably imply that agents who desire one of these distinct goods, without desiring the others, are at least partially praiseworthy for actions that reflect this desire.[26] The problem cases discussed in this chapter do *not* require a pluralistic normative theory in order to arise. For even a monistic normative theory, which identifies a single consideration as good, can admit of multiple interpretations as to what *counts* as a desire for the good. The first problem case below illustrates how a monistic, Kantian normative theory can give rise to the ambiguity.

> **Torture:** Elaine is a CIA agent, and is ordered by her superiors to torture a prisoner; she disobeys orders and refuses to do so, sacrificing her career. Elaine is a utilitarian, and believes that actions which fail to maximize utility are wrong; she concludes that torturing the prisoner would be wrong because it would cause more pain than it would prevent. Elaine's decision is motivated by three desires: the desire to act rightly, the desire to maximize utility, and the desire to avoid causing pain; she believes that all three desires can be satisfied simultaneously by refusing to torture the prisoner.
>
> As it turns out, torturing the prisoner *would* be wrong, but not for quite the reasons that Elaine thinks: The correct normative theory is a Kantian one, on which actions are wrong when and because they constitute treating an agent as a mere means. The nature of pain is such that to inflict it on a person always constitutes treating him as a mere means. So the painfulness of torturing the prisoner does provide a reason not to do it, although this reason has nothing to do with utility.

Is Elaine praiseworthy for her decision not to torture the prisoner? She does act rightly, and her action does reflect a desire for the right *de dicto*; but, on AG accounts, these factors are irrelevant to moral worth.[27] The real question is whether or not Elaine's action reflects a desire

---

26 On Arpaly and Schroeder's view, this would seem to be a fairly clear case of an agent desiring a partial good.

27 Or, at least, relatively insignificant. In the previous chapter I conceded that an AG account could allow that *some* aspects of virtue do not consist in correct responsiveness to the actual good and bad; but these must amount to only a small part of virtue, I argued, in order for the motivations of AG accounts to be preserved. I assume from this point forward that *all* of virtue consists in responsiveness to the actual good and bad. But readers who prefer to allow that some small part of virtue does not are free to do so. In cases in which I assert

for the actual good, and it is not obvious how this question should be answered. On the one hand, Elaine is *not* motivated by a desire to refrain from treating the prisoner as a mere means. Since, as stipulated, the correct normative theory is concerned only with whether we treat persons as mere means, we might conclude that Elaine does not desire the actual good. Perhaps Elaine is just "lucky" in that her morally-irrelevant worries about utility cause her to act in accord with the correct, Kantian theory; if this analysis is correct, then Elaine is not praiseworthy.

On the other hand, Elaine *does* want to refrain from causing the prisoner pain. And the painfulness of torture *is* part of the story of why torture is actually morally bad – the painfulness of torture is what makes it the case that it constitutes treating someone as a mere means. Furthermore, the badness of pain is not a coincidence – we have stipulated that the nature of pain is such that, when inflicted intentionally, it always makes an action wrong, by making it an instance of treating a person as a mere means. So Elaine is motivated by a desire to avoid a feature that non-accidentally makes actions wrong. Does this count as a desire for the good (or, at least, an aversion to the bad)? If so, then Elaine *is* praiseworthy.

Cases like Torture are likely to occur in the real world. There is widespread disagreement about which normative theory is correct, which implies that many agents are mistaken; even so, many of these mistaken agents act rightly, and do so in a way that seems to track actual goodness non-accidentally. And although we may have a full account of the motivations of these agents, as well as of the correct normative theory, it is not clear whether AG accounts imply that these agents are praiseworthy or not. The answer to this question depends on what counts as a desire for the actual good. On one plausible understanding, agents like Elaine desire the actual good; on

---

that an agent is not praiseworthy, these readers can understand the claim to be that the agent is at best only slightly praiseworthy.

another plausible understanding, they fail to.

There are also problem cases that arise from ambiguities surrounding what counts as indifference towards the actual bad:

> **Psychopathy:** Newman is a psychopath. His general intelligence is higher than average, and he has a particularly good understanding of the psychology of others, which allows him to manipulate people very effectively. Newman has just perpetrated a financial scam, accepting large "investments" under false pretenses and then absconding with the money. Newman's action was motivated solely by a desire to enrich himself; he was aware that doing so would cause harm to others.
>
> Although Newman's childhood therapist lectured him on why it is wrong to harm others, he did not (and does not) understand how the harmfulness of an action could provide reasons for him; nor does he understand that other people have rights, or why these rights should factor into his own deliberations. As it turns out, Newman's action *is* wrong. Individuals have a number of rights, including the right not to be harmed; the correct normative theory states that any action which violates the rights of others is wrong.

Psychopathy is a condition characterized by a lack of concern for the rights and welfare of others, poor impulse control, and repeated criminal behavior. Many psychopaths, like Newman, have average or above-average intelligence, and are capable of perpetrating complex frauds; also like Newman, psychopaths are generally unmotivated by the considerations that seem morally important to normal agents.[28] The fact that psychopathy is apparently an innate condition, coupled with the fact that psychopaths seem to be in some sense *unable* to appreciate their moral reasons, has led some philosophers to argue that psychopaths cannot be blameworthy for their actions.[29] Others have defended a contrary position, appealing to the fact that psychopaths are both in control of their actions and aware of the features that make their actions wrong.[30]

---

28 See Cleckley (1964), Hare (1993), Kiehl (2008), and Scott (2014) for general descriptions of the condition.
29 See, e.g. Benn (2000), Levy (2007), Litton (2008), Haji (2010), Shoemaker (2011), and Nelkin (2015).
30 See, e.g. Greenspan (2003), Maibom (2005, 2008), and Talbert (2008, 2012, 2014).

Because psychopaths commit a disproportionate number of crimes[31], the question of whether they are blameworthy for their bad actions is of particular importance. If their condition *does* excuse them from blame, it may be unjust to punish them. Since the law generally does not allow defendants to be excused on the basis of psychopathy, our current judicial practices may need to be dramatically revised.[32] Unfortunately, as with the previous case, existing AG accounts do not provide a clear judgment as to whether psychopaths like Newman can be blameworthy.

AG accounts are sometimes interpreted as implying that psychopaths *are* blameworthy.[33] Because of their intelligence and their generally good understanding of the world, psychopaths are typically aware of the harms that they inflict on others. Newman, for example, knows that he harms his victims when he defrauds them of their property. Since this feature makes the action wrong, and Newman is not deterred by his knowledge of this feature, there is clearly *some* sense in which his action reflects indifference towards that which is actually bad. Thus, it may seem as though AG accounts commit us to the conclusion that psychopaths like Newman are blameworthy.

However, this is not the only possible conclusion, for there is much that Newman does *not* know: He does not appreciate that other people have rights, nor understand that these rights should factor into his own decision-making. These facts are part of the full explanation for why Newman's action is wrong. And one might think that, since Newman is essentially ignorant of the fact that others have rights, he cannot express a lack of concern for the rights of others by

31 Hare (1993) estimates that there are only a few million psychopaths in North America, but that these psychopaths commit more than half of all serious crimes.(p.74, 87).

32 See Lyon and Ogloff (2000) for a discussion of psychopathy and the law. Interestingly, some statutes, as well as the Model Penal Code, have apparently been *designed* so as to exclude psychopaths from qualifying for an insanity defense. See Model Penal Code §4.01(2), which excludes defendants from claiming insanity on the basis of "an abnormality manifested only by repeated criminal or other antisocial conduct."

33 E.g. Levy (2007) attributes this implication to AG accounts.

acting. By way of analogy, suppose that, unbeknownst to humans, trees have rich inner lives which end painfully when they are cut down. While this presumably gives us an objective moral reason not to cut down trees, it does not imply that lumberjacks are blameworthy for doing so. Presumably, their lack of awareness of the inner lives of trees excuses them from blame – their actions cannot express a lack of concern for the well-being of the trees, because they are not aware that trees *have* any well-being.[34] Similarly, because Newman is unaware that others *have* rights, it may be impossible for his actions to reflect a lack of desire that these rights be respected. There is a sense, therefore, in which Newman's cognitive limitations apparently prevent his actions from reflecting indifference towards the actual bad.

Recall that on AG accounts, non-moral ignorance can exculpate while moral ignorance cannot. It may therefore seem as though the problem posed by psychopaths can be resolved by asking whether their cognitive abnormalities lead to ignorance that is moral or non-moral in character. An answer to this question would, in fact, resolve the ambiguity; the problem is that this question is a difficult and substantive one. By virtue of his general and social intelligence, Newman has a detailed and accurate understanding of other human beings as entities with mental lives and capacities much like his own. What he does *not* know, by stipulation, is that these other persons have *rights*. Does this amount to non-moral ignorance, like that of the lumberjack who is unaware that trees have inner lives? Or does it amount to a form of moral ignorance? Might, for instance, not knowing that others have rights simply be equivalent to not knowing that it is wrong or bad to treat others in certain ways?

Ultimately, I will argue, the ambiguity surrounding psychopaths is best addressed by

34 A somewhat similar thought experiment using aliens and grass is employed by Levy (2007) and Shoemaker (2011) to illustrate a point about the moral status of psychopaths. I discuss this example at length in Chapter Six.

investigating a different question. Newman knows that he harms his victims, and thus his action reflects indifference to causing harm. But Newman does *not* know that his victims have rights – or even understand what rights *are* – and thus his action *cannot* reflect indifference to the violation of rights. To evaluate Newman, I claim, we will need to know what would *count* as indifference towards the actual bad in this case – is indifference towards harm sufficient, or would indifference towards rights be required?

I conclude this section by discussing a third problem case:

> **Clinic Bomber:** George is deeply concerned with the well-being of persons, and strongly desires to prevent persons from being killed. Because he believes that fetuses are persons, he believes that he can save persons by preventing abortions. Accordingly, he places a small bomb in an abortion clinic and detonates it at a time when he knows the clinic will be unoccupied. The resulting damage to the facility, which is located in an area with limited access to abortion, forces it to close for several weeks and prevents a number of abortions which would otherwise have taken place.
>
> George originally became convinced that fetuses were persons when he read a description of their biology. A fetus has a complete and unique human genome, and this, George thinks, endows it with personhood. This belief is false – a being actually requires certain psychological properties in order to be a person – but George came to acquire it by reasoning responsibly and without self-deception. Fetuses do not in fact have the required psychological properties, so George's action does not actually save any persons. Even so, saving persons *is* morally good, and, if fetuses *were* persons, George's action would have been morally right.

Like the previous cases, I take Clinic Bomber to be realistic. A survey of the popular rhetoric surrounding abortion demonstrates that opponents sometimes appeal to the fetus's genes as a reason to think it is a person[35]; some agents, we may safely infer, must be convinced by such appeals. Since agents sometimes do bomb abortion clinics – and some of these agents may have beliefs like George's – we should want an account that allows us to evaluate them. What should AG accounts say about the moral worth of George's action?

35 E.g. Terzo (2013). I return to this topic in Chapter Five.

On the one hand, George's action reflects a desire for something which is actually good – he wants to save persons from being killed. Of course, we have stipulated that he is mistaken about whether fetuses are persons, and so he does not actually save any persons when he acts. But mistakes of this kind, it might be argued, generally do not prevent agents from being praiseworthy. Imagine a rescuer who goes to heroic lengths to recover a life raft that she believes to be occupied, but which in fact turns out to be empty. This agent fails to save any lives; but her action nevertheless reflects a strong *desire* to save human life, and she seems to be praiseworthy. We might analyze George's action similarly – since it reflects a strong desire to save people from death, it reflects a strong desire for the actual good. On this analysis, AG accounts will tell us that George is praiseworthy.

On the other hand, the reason that George thinks fetuses are persons is because of their genetic properties; and, as stipulated, having a complete human genome does *not* make a being a person. Actual persons, such as adult humans, possess certain psychological properties that *make* them persons. George does not attribute any of these properties to the fetuses which he tries to save. So while George's action *does* reflect a desire to save people, it does *not* reflect a desire to preserve any of those features that actually *make* persons persons. Instead, it reflects a desire to save things-with-complete-human-genomes. As stipulated, saving things-with-complete-human-genomes is not morally important. So, we might conclude, George's action actually reflects a desire for something that is morally irrelevant, and he is not praiseworthy.

Like the previous case, the ambiguity here can be interpreted as one concerning moral versus non-moral ignorance. Does George's false belief that fetuses are persons represent non-moral ignorance, analogous to the false belief of a would-be rescuer who believes that there are persons in an empty lifeboat? Or is it really a variety of moral ignorance, reflecting confusion as

to which kinds of properties make beings morally valuable? As in the previous case, the ambiguity can be most fruitfully addressed by investigating a different question: What counts as desiring the actual good? If desiring that persons be saved counts, then George is praiseworthy. But perhaps something else is required – perhaps only a desire to save those beings with the actual personhood-conferring properties counts. If so, then George's action does *not* reflect a desire for the actual good, and George is not praiseworthy.

My aim in presenting these problem cases is to show that, even when the correct normative theory and the attitudes reflected in an agent's action are stipulated, it may be unclear whether an action reflects a desire for the actual good. We require something *more* in order to determine whether these agents are praiseworthy or blameworthy; the "something more" that we require, I argue shortly, is a procedure for determining which desires count as desires for the actual good.

It is worth reiterating at this point that the problem described here is not unique to the particular account provided by Arpaly and Schroeder – other existing AG accounts fail to specify *precisely which* attitudes count as appropriate responses to morally-important considerations. The problem also arises, for instance, on Markovits's (2010, 2012) account of praiseworthiness. According to Markovits, agents are praiseworthy when they are motivated by the moral reasons that actually justify their actions. Here, the problem can be described as an ambiguity with respect to precisely which reasons justify actions. An action that saves persons is right, in the world of Clinic Bomber. But is such an action justified by a reason to save persons as such? Or is it justified by a reason to save those beings with the personhood-conferring psychological properties? Since George is motivated by the former reason but not the latter, we need an answer to this question in order to judge the moral worth of his action.

Fortunately, I think that there *is* a principled way to answer these questions. Before laying the groundwork for my proposed solution, however, I revisit two features of Arpaly and Schroeder's account – the distinction between complete and partial goods, and the requirement that goods be desired under the correct conceptualization – and examine whether these provide an alternative means of eliminating the ambiguity.

II. Partial Goods and the Correct Conceptualization of the Good

Recall that Arpaly and Schroeder distinguish between complete and partial goods, and claim that agents can be praiseworthy for actions that reflect a desire for either. The *complete good* is meant to consist in the entirety of that which is identified as good by the correct normative theory. So, assuming hedonistic utilitarianism to be true, to desire the complete good would be to desire that the balance of pleasure over pain be maximized. A *partial good* is anything which, according to the correct normative theory, we have a *pro tanto* moral reason to do or to bring about. According to hedonistic utilitarianism, we have a *pro tanto* moral reason to increase the pleasure or alleviate the pain of any particular person, so an agent would desire the partial good by desiring to increase the pleasure or diminish the pain of some individual.

The distinction between complete and partial goods makes it possible for this account to judge agents praiseworthy even though they fail to desire the complete good. Arpaly and Schroeder motivate the distinction by appealing to the fact that agents in centuries past often seem to have been praiseworthy even though they lacked desires for the complete good.[36] We might imagine an agent who, for instance, desires that the rights of some subset of people be respected, while lacking a corresponding desire about the rights of some other subset (say, women). Assuming that the correct normative theory commands us to respect the rights of all

36 pp.165-6, 194-5.

people, this agent does not desire the whole good. But it does seem plausible to describe him as desiring a *part* of it – after all, he does have a *pro tanto* moral reason to respect the rights of the subset of people that he acknowledges as having rights – and it also seems plausible to conclude that he is praiseworthy for actions that reflect this desire.

At issue here is whether the partial good can also be invoked to provide unambiguous attributions of moral worth in the problem cases described above. Consider how this strategy might work in Torture. The complete good, as stipulated in this case, consists in no person's being treated as a mere means, which Elaine does not desire. However, Elaine *does* desire that she not cause pain to the prisoner. Whether or not Elaine is praiseworthy will depend on whether or not the desire to avoid inflicting pain counts as a partial good, given the stipulated normative theory. If we resolve this question, we will know whether or not Elaine is praiseworthy.[37]

It seems to me that a plausible case can be made for either answer, depending on precisely how we understand the relationship between complete and partial goods. On the one hand, as stipulated, we always treat someone as a mere means when we inflict pain. Therefore, refraining from inflicting pain is *part of* not treating people as mere means; on a fairly intuitive understanding of the term "partial," it seems that Elaine's desire *does* count as a desire for the partial good. On the other hand, it is not clear that Elaine has a *pro tanto* moral reason not to inflict pain – perhaps the reason not to inflict pain is better described as an *instrumental* one, since inflicting pain is only bad by virtue of the further fact that it constitutes treating someone as

37 Arpaly and Schroeder describe a case similar to Torture, and discuss the possibility that an appeal to partial goods will allow for an attribution of praiseworthiness. It is unclear whether they intend for their account to allow for praiseworthiness in Torture as I have described it. They appeal to the possible truth of a pluralistic normative theory, on which preventing pain is one of many goods, whereas I have stipulated a monistic Kantian theory in Torture. They also appeal to the fact that real-world agents are unlikely to be wholeheartedly devoted to particular moral theories, meaning that many such agents will desire the actual good even if their explicit moral beliefs are false. But in my description of the case I have supposed that Elaine is a wholehearted utilitarian. (2014a) pp.198-199, (2014b). See also Hurka (2014) for discussion of a similar case, which I return to in Chapter Three.

a mere means. If we keep to the definition of partial goods offered by Arpaly and Schroeder – as goods which we have a *pro tanto* reason to pursue – it may turn out that Elaine does not desire a partial good. In any case, however, this uncertainty results from different understandings of what a partial good consists in, and does not reflect a deep ambiguity in Arpaly and Schroeder's account. Whichever interpretation we prefer, the account will provide an answer as to whether Elaine is praiseworthy.

However, even if this distinction between complete and partial goods is adequate to allow unambiguous judgments of moral worth in cases like Torture, it seems inadequate to allow unambiguous judgments in others. Return to Clinic Bomber. George wants to save persons from death; since (we may suppose) this is the entirety of what is commanded by the correct normative theory, it seems that if George is to be understood as desiring *any* good, it must be the complete good rather than a partial one. But the reason the case is puzzling is that it is unclear whether we *should* understand George as desiring anything good. The worry here is that a desire to protect persons *qua* persons may not represent good will *at all* and thus may be unsuitable for grounding praiseworthiness.

The preceding discussion may suggest another strategy for dealing with the ambiguity – it may seem that cases like Clinic Bomber can be best addressed by asking which *conceptualization* of the good is the correct one. We might understand the two desires at issue in this case – the desire to save persons *qua* persons, and the desire to save beings with the actual personhood-conferring properties – as desires for the good under different conceptualizations. If we can determine which conceptualization of the good is the correct one, we can determine whether George's desire represents good will or not.

One worry concerning this strategy is that it is not clear that these really *are* two different

conceptualizations of the same desire. Although the set of persons might be coextensive with (or even *necessarily* coextensive with) the set of beings with the personhood-conferring properties, it is not obvious that *being a person* is identical to *having psychological properties X, Y, and Z.* This dissertation is not the venue for a discussion of precisely what is required for two properties to be identical to one another, but it seems at least possible that the property of personhood is distinct from the property of having the actual personhood-conferring features. If so, the appeal to conceptualization is not applicable to Clinic Bomber. Conceptualization is even less likely to be relevant to Torture and Psychopathy: It seems clear that the property of inflicting pain without consent is not *the same as* the property of using someone as a mere means, and that the property of causing harm is not *the same as* the property of violating a right.

But suppose that being a person *is* identical to having psychological properties X, Y, and Z. The strategy at issue here would have us identify which conceptualization of this single feature is the "correct" one. How might we accomplish that? Arpaly and Schroeder's view is that the correct conceptualization of the good is identified by the correct normative theory – whichever concepts the correct theory uses are the concepts under which an agent must desire the good in order to be praiseworthy.[38] Which concepts does the correct normative theory use? In my description of Clinic Bomber, I stipulated that it was right to save persons. Perhaps this implies that {personhood} is the relevant concept. So, to be praiseworthy, George's action must reflect the desire {that persons be saved.}

George's action does reflect this desire, so this interpretation would apparently imply that he is praiseworthy. As noted, however, there is some intuitive reason to think that George is *not* praiseworthy; it may not seem like his desire {that persons be saved} reflects good will, given

38 2014a, p.15,pp.176-8.

that he does not know what persons *are*. Perhaps we are *wrong* about which conceptualization the normative theory uses. Perhaps it really commands us to save persons under their description as beings with properties X, Y, and Z; if we describe the normative theory as commanding us to save "persons", this could be because we use "person" as a shorthand for a being with the relevant properties, or even because the true "meaning" of the normative theory is somehow hidden from casual observers.

The normative theories used in these examples were merely assumed for the sake of illustration; in these hypothetical cases, we are of course free to stipulate what the correct normative theory is as well as how it should be interpreted. The preceding discussion is intended to illustrate the fact that, if we want to rely on the appeal to conceptualization to resolve cases like Clinic Bomber, we will have two options. One is to maintain that a given normative theory really does identify the good and bad under a particular conceptualization, but that the correct conceptualization is not trivial to discover. Many normative theories include prohibitions against treating persons in certain ways; in our standard way of articulating these theories, we generally employ the concept of {personhood}. If we take this first option, we will need to ask whether this standard articulation is correct or not; we must consider the possibility that we are mistaken about the concepts which our normative theories employ. Kant, for instance, tells us that we are required to treat persons as ends in themselves. But perhaps what he really means – or what he *should* mean – is that we are required to treat beings with properties X, Y, and Z as ends in themselves.

The upshot will be that the development and exegesis of normative theories may be much more difficult than we previously believed. In addition to determining *what* is right and wrong, to fully describe the correct theory will require us to determine the correct *conceptualization* of

what is right and wrong. We might agree that it is right to save certain kinds of beings, but there will be an open question as to *precisely* why it is right to do so – is it in virtue of their personhood, or in virtue of the properties that make them persons? In the context of a thought experiment, we can stipulate that it is one or the other. But if we are interested in assessing moral worth in the actual world, we will need to know which conceptualization the *true* normative theory uses, and it is not obvious how we could make this determination. Our intuitions about which actions are right and wrong will certainly not help, as actions that affect persons are coextensive with those that affect beings with the personhood-conferring properties.

The second option is to deny that a given normative theory identifies the good or bad under a single, correct conceptualization. This does not amount to giving up on the problem cases, but it does require us to abandon the idea that the correct normative theory will tell us all we need to know to determine moral worth. The theory will tell us what is good and bad, but it will not tell us which conceptualizations of the good and bad are relevant to character and moral worth; to identify the relevant conceptualizations, and to resolve the problem cases, we will require a further, substantive story.

As noted earlier, I do not think that the features described in these problem cases *should* be understood as different conceptualizations of the same feature, and so I reject the appeal to conceptualization at the outset. But if we *do* think that the appeal to conceptualization is the best strategy, then I believe that we should take the second option. The first option would, in my view, amount to taking a rather surprising view of normative theories; it requires us to conclude that they contain additional information over and above a complete account of what is right and wrong.[39] Furthermore, as noted, it is difficult to see how we could identify the "correct"

39 It might be objected that, irrespective of the problem cases described here, we have independent reason to

conceptualization of the good for a non-stipulated normative theory, and thus it seems unlikely that we would be able to make progress on problem cases that arise in the real world. The second option, in contrast, *does* offer the prospect of progress. If in fact there is no "correct" conceptualization identified by the normative theory, we are faced instead with the question of which of multiple conceptualizations is *relevant* to moral worth. In the next chapter, I discuss the question of which *features* are relevant to moral worth, and in Chapter Four I describe how we can go about answering this question. But the model I describe is compatible, with a few minor changes, with the claim that the putative "features" merely represent different conceptualizations. Readers who prefer the appeal to conceptualization are free to continue thinking of them as such; as I discuss at the end of Chapter Four, the procedure that I use for determining which features are relevant could, *mutatis mutandis*, also allow us to determine which conceptualizations are relevant.

think that normative theories identify the good under a particular conceptualization. After all, how else would we have grounds to exclude desires for the good *de dicto*, as well as obviously irrelevant attitudes such as a desire to promote certain neural states? In the next chapter, I offer an explanation for how these attitudes can be excluded even if the normative theory lacks any information on the correct conceptualization of the good.

Chapter Three

The Relevant Right-Making Features

I. Right- and Wrong-Making Features

Before proceeding, it will be helpful to explicitly introduce the concepts of *right*- and *wrong-making* features. AG accounts hold that the moral worth of an action depends on the extent to which it reflects the correct attitudes towards certain kinds of considerations; the considerations that are relevant are those that actually determine the deontic status of actions. This central principle of AG accounts can be articulated in various ways; Arpaly and Schroeder (2014a), whose characterization I have been borrowing, describe moral worth as depending on whether an action reflects a desire for the actual good or actual bad. The central principle can alternatively be described as the claim that moral worth depends on whether an agent *responds correctly* to the actual *right*- or *wrong-making features* of actions. Right-making features are those which actually make actions right (or which make it so that we have *pro tanto* moral reasons to perform them) and wrong-making features are those which make actions wrong (or which make it so that we have *pro tanto* moral reasons not to perform them). To respond correctly to these features is to have pro-attitudes towards the right-making features and to have anti-attitudes towards the wrong-making features.

Arpaly's (2003) account appeals to an agent's responsiveness to right- or wrong-making features; Arpaly and Schroeder's (2014a) largely abandons this terminology in favor of desires for the actual good or bad. For present purposes, these two articulations of the central principle can be understood as equivalent – one desires the actual good and abhors the actual bad iff one is properly responsive to the features that make actions right and wrong. The change in terminology

in this section is not intended to represent a substantive revision to the content of any existing AG account, but rather to allow the problem facing these accounts to be described more clearly.

Just as the actual good and bad depend on which normative theory is correct, so do the right- and wrong-making features of actions. The fact that an action involves treating an agent as a mere means, for example, might be a wrong-making feature in a Kantian universe but not in a utilitarian one. And, just as is the case for the actual good and bad, it may be unclear which features of an action are right- or wrong-making, even when the correct normative theory is held fixed. In Clinic Bomber, for instance, should we count the fact that an action saves persons as a right-making feature? Or the fact that an action saves a being with certain psychological properties? Or both? Or neither? As before, the correct answer will affect the moral worth of George's action. If saving persons is a right-making feature, he will be praiseworthy, as his action shows correct responsiveness to this feature; if not, then he will not be praiseworthy.

When we inquire as to the right- or wrong-making features of an action, we are asking for the features that explain why this action has a certain property – the property of being right or wrong. And there are often multiple features that explain why a particular object has a particular property. By way of analogy, consider the following question: What is it about red wine that makes it healthy to consume in moderation? That is, what is the healthy-making feature of red wine?

One answer: Red wine is healthy because it reduces one's risk of having a heart attack.

Another answer: Red wine is healthy because it has anti-oxidant properties.

And another: Red wine is healthy because it contains tannins.

A few preliminary remarks: First, I take it that each of these is a correct and informative[40]

40 I take the following answer: “Red wine is healthy because it is healthy”, to be correct but *un*informative.

answer to the question “Why is red wine healthy?” Second, I take it that each of these answers describes a feature *of wine* as opposed to some other fact about the universe. An answer such as “Red wine is healthy because the human vascular system responds well to tannins” might be correct and informative, but would describe a feature of humans rather than of wine, and is therefore excluded from this list. Third, while there is clearly some relationship between the features described in these answers, they represent distinct features rather than different conceptualizations of the same feature; e.g., having anti-oxidant properties is not the same feature as containing tannins. Instead, the relationship between the features cited here seems to be better described as a kind of (possibly partial) *in-virtue-of* or *making-the-case* relationship: the presence of the features lower in the list, taken in conjunction with certain background facts about the world, *makes it the case that* red wine has the higher features. So the fact that wine contains tannins, in conjunction with the fact that tannins are anti-oxidants, makes it the case that wine has anti-oxidant properties. This feature, in conjunction with the fact that anti-oxidant properties prevent heart attacks, makes it the case that wine can prevent heart attacks. And this feature, finally, in conjunction with the fact that heart attacks are injurious to one's health, makes it the case that wine is healthy.

The main observation I wish to make, however, is this: *All* of these features are genuinely features that make wine healthy, and we have no basis for picking out any particular feature as the “real” healthy-making feature. It does seem that certain features will be more important than others in certain contexts. A vintner who wants to produce the healthiest possible wine will focus on making wine that has high levels of tannins. A pharmaceutical chemist who wants to develop a pill with the same health effects as wine will be most interested in wine's anti-oxidant properties. And a physician who is trying to decide whether to advise his patient to drink wine

will be most interested in the fact that wine is beneficial for the heart (as opposed to, say, the liver). So, when we ask what makes wine healthy, the context of our inquiry or our reasons for asking may point to a particular feature as the one of interest. But there seems to be no basis for claiming that any particular feature is *the* feature that makes wine healthy. Thus, if we had some theory that required us to input *the* healthy-making feature of wine, we would be stuck. There is no such singular feature, and we would forced to choose which feature to input from an array of possibilities.

The reason that existing AG accounts are not complete is that any plausible normative theory picks out a list of multiple features that make actions right or wrong. Return to the case of Torture. As stipulated, the correct normative theory is a Kantian one on which actions are wrong when they constitute treating a person as a mere means. What is it about torturing the prisoner that would make it wrong for Elaine to do so?

One answer: Torturing the prisoner would constitute treating him as a mere means.

Another: Torturing the prisoner would cause him pain.

And another: Torturing the prisoner would cause certain events to occur in his brain.

Each of these features represents a correct and informative answer to the question "Why would it be wrong for Elaine to torture the prisoner?" Each is a feature of the action itself, rather than a feature of some other part of the world or a background condition. And, although the features are not identical to one another[41], they are obviously *related* – some features of the action, in conjunction with certain additional facts, make it the case that the action has other features. The fact that the action causes certain neural events to occur, in conjunction with

41 *Some* of the features, at least, are non-identical. On certain views of mind, neural events *are* identical to phenomenal events like painful sensations, so it is possible that the two "lowest" features on this list are identical to one another. This possibility does not significantly impact the discussion in this chapter, but at the end of Chapter Four, I pause to address some potential worries related to questions of conceptualization.

whatever psychophysical laws or principles connect these events to conscious experiences, makes it the case that the action causes pain. The fact that the action causes pain, in conjunction with the fact that intentionally inflicting pain constitutes treating a person as a mere means, makes it the case that the action constitutes treating the prisoner as a mere means. And the fact that the action constitutes treating the prisoner as a mere means, in conjunction with the truth of the stipulated Kantian theory, makes it the case that the action is wrong.

On my analysis, existing AG accounts are not complete because they require us to input *the* right- or wrong-making features of actions in order to produce a judgment of moral worth. The problem is that, just as there is no such thing as *the* feature that makes red wine healthy, there is no such thing as *the* feature that makes an action right or wrong. Each of the features described above is a genuine feature that would make torturing the prisoner wrong, and we have no basis on which to privilege one of them as the "real" wrong-making feature. As such, it is not clear which feature to input into our formula for moral worth, and the formula will, in some cases, produce different answers based on which feature we choose. Fortunately, the analogy to the healthy-making features of wine also provides a clue as to how AG accounts can be made complete. For, while none of the wrong-making features listed above is more genuine than the others, some right- or wrong-making features are *relevant in certain contexts*, while others are not. The solution lies in identifying which kinds of right- and wrong-making features are relevant in the context of attributing moral worth; once we identify the set of *relevant* right- and wrong-making features, an AG account should be able to provide unambiguous attributions of moral worth based on an agent's responsiveness to that restricted set.

More precisely, I propose the following:

For each normative theory N, there is some subset S of the features that make

> actions right or wrong, such that an agent's character depends only on whether he has the appropriate attitudes towards the features in S, and such that the moral worth of an agent's actions depends only on whether or not they reflect the appropriate attitudes towards the features in S.[42]

To reiterate: My suggestion is *not* that certain features identified by the correct normative theory are "real" right- or wrong-making features, at the expense of all others. Rather, it is that moral worth *depends on* an agent's responsiveness to some of the right- or wrong-making features but not to others. Only an agent's responsiveness to the features within this limited subset "counts", at least *for the purposes of evaluating character and attributing moral worth*. (To return to the terminology used in earlier chapters, only a desire for a feature within this limited subset *counts as a desire for the actual good*.) If an agent's action reflects the correct attitudes towards the features in S – such as a desire for the right-making features or an aversion towards the wrong-making ones – the agent is praiseworthy. If an agent's action reflects the wrong attitudes towards the features in S – such as a desire for the wrong-making features, or indifference towards the right-making ones – the agent is blameworthy. If we have a reliable procedure for determining which features belong in S, this addition to AG accounts should make them *complete*, and should allow them to make unambiguous attributions of moral worth in the full range of previously problematic cases.

Return, for instance, to Clinic Bomber. George's action reflects a desire for one right-making feature – the desire to save persons from being killed. But George's action fails to reflect

---

42 One advantage of the appeal to conceptualization, as discussed in the last chapter, is that it allowed us to exclude attitudes that were clearly irrelevant (such as those towards neural states) as well as attitudes towards the good *de dicto*. My proposed solution easily accomplishes the first goal – I treat an action's effects on neural states as a distinct feature, and, as I will argue in the next chapter, this feature does not belong in S. The exclusion of attitudes towards the good *de dicto* can also be accomplished if we treat the rightness of an action itself as a vacuously right-making feature – e.g. if we allow that "This action is right" is one feature that explains its rightness. This vacuously right-making feature can also be excluded from S, and thus we can explain why attitudes towards the good *de dicto* are irrelevant to moral worth without appealing to conceptualization.

a desire for another right-making feature – it does not reflect the desire to save beings with the particular psychological properties that actually confer personhood. Whether or not George is praiseworthy depends on which of these features fall within the relevant subset S. If the fact that an action saves persons falls within S, then George is praiseworthy – his action reflects responsiveness to a right-making feature of the appropriate kind, one which "counts" as a desire for the actual good. If, on the other hand, the fact that an action saves a person does not fall within S, then George's action does not reflect responsiveness to the right-making features of the relevant kind, and thus he is not praiseworthy. It might turn out, for instance, that the sole right-making feature in S is the fact that an action saves a being with personhood-conferring psychological properties, a feature to which George is indifferent. It is also conceivable that *neither* of these features is contained in S, in which case George would not be praiseworthy, or that *both* features are contained in S, in which case George would presumably be partially praiseworthy for responding to one feature but not to the another.

The success of the solution proposed here ultimately depends on our ability to properly restrict the set of relevant desires; it requires some procedure for determining, given the correct normative theory, which right- and wrong-making features belong in the relevant subset S. I will offer such a procedure in the next chapter. The remainder of this chapter is devoted to two remaining preliminary tasks. In the next section, I briefly discuss the relationship of my proposal – on which there are multiple right-making features at different levels – to several existing treatments of moral reasons and morally-relevant features. In the final section, I address two possible objections to my proposal.

II. Morality and Multiple Levels

I have argued that a given normative theory identifies a range of right- and wrong-making features, and that moral virtue and moral worth depend only on an agent's responsiveness to a subset of these features. In the next chapter, I will argue that these features can be meaningfully grouped according to their "level" – feature X is at a lower level than feature Y if Y is present (at least partly) in virtue of X but not vice-versa. But before describing the multi-level structure that I propose, it will be helpful to discuss some existing views which also group moral reasons or morally-significant features into different levels. I focus on two such views here – one defended at length by Daniel Star (2011, 2015), and the other discussed more briefly by Julia Markovits (2010) and Thomas Hurka (2014). Star's view, I argue, is only superficially similar to mine; Markovits's and Hurka's views, in contrast, may partially prefigure the approach defended in this dissertation.

Star offers his account in an effort to reconcile two plausible yet seemingly inconsistent propositions. The first is that ordinary people, who have little or no knowledge of normative theory, can act virtuously; the second is that moral philosophers who investigate normative theories are not wasting their time. These two propositions may appear to be in conflict, because the normative theories developed by philosophers are generally unknown to the folk. If knowledge of the correct normative theory meaningfully contributes to one's knowledge of the good, it may be unclear how the folk can have enough knowledge to act virtuously. If knowledge of the correct normative theory does *not* meaningfully contribute to one's knowledge of the good, it may seem that moral philosophers are wasting their time when they search for this theory.

Star's way of reconciling these propositions requires him to distinguish between *fundamental* moral reasons, which are identified by the correct normative theory, and *derivative*

moral reasons, which are the kinds of reasons that virtuous folk respond to. Though the folk cannot respond *directly* to the fundamental reasons, they can do so *indirectly* by responding to the derivative reasons. On Star's view, reasons are supposed to be a kind of *evidence* – I have a reason to act in a certain way just in case I have *evidence* that I should act in that way. Responsiveness to derivative moral reasons is supposed to be morally meaningful because these derivative reasons represent genuine evidence that an agent ought or ought not to act in certain ways. The fact that an action causes pain is evidence that one ought not to perform it, and thus there is a genuine moral reason not to cause pain; responsiveness to these kinds of reasons is, according to Star, sufficient for virtue.

It is important to distinguish my view from Star's, because the two might at first seem to be similar. One might think that the problem cases from the previous chapter can be understood as concerning responsiveness to derivative reasons rather than direct responsiveness to fundamental reasons. In Torture, for instance, Elaine responds properly to what seems to be a derivative reason not to torture the prisoner, without responding *directly* to the fundamental reason. And although I have not yet offered an account of which attitudes are relevant to virtue and moral worth, I will ultimately argue that it is attitudes towards features like the painfulness of an action that count. Thus, on my view as well as Star's, it will turn out that ordinary agents like Elaine can respond in the ways required for them to be virtuous.

But the similarities between my account and Star's are largely superficial. First, I do not endorse Star's view that reasons are a form of evidence, which I take to be fairly central to his account of derivative and fundamental reasons.[43] On my view, the relationship between the

43 I do not endorse *any* account of what reasons are, as the account of moral worth I ultimately develop does not require one.

different right- and wrong-making features of a given action is a form of in-virtue-of or making-the-case relationship; Star, in contrast, views the relationship between fundamental and derivative reasons as an epistemic one, with the former serving as evidence of the latter. Second, the distinction between fundamental and derivative reasons cannot be usefully applied to all of my problem cases, in particular Clinic Bomber – the relationship between the reason to save persons and the reason to save beings with personhood-conferring properties does not seem to be one between a derivative and a fundamental reason. Finally, Star's overall account of virtue is much more forgiving than the one that I ultimately defend, as he is willing to describe as virtuous any agent who "does her best to respond to reasons."[44] In contrast, I hold any agent vicious who fails to respond properly to the right- and wrong-making features in S; a major implication of this view, defended in Chapters Five and Seven, is that the folk are often much *less* virtuous than we might previously have believed.

I turn now to an alternative schema, which is presented in much less detail but which is potentially more similar to the view I defend here. Hurka (2014) suggests in passing that agents might be praiseworthy in virtue of their attitudes towards *derivative* duties; this would allow that agents could be praiseworthy even if they wholeheartedly endorsed the wrong normative theory.[45] Hurka's suggestion seems to be aimed at cases somewhat like Torture. The idea is that an *ultimat*e duty, such as the duty not to treat others as mere means, might produce a number of derivative duties, such as the duty not to inflict pain on others. Because an agent like Elaine does respond correctly to her derivative duties, she does desire the good on some level, and therefore displays virtue.

44 2015, p.xi.
45 p.502.

Markovits (2010) similarly suggests that a given normative theory, while it may only identify a limited range of considerations that are *fundamentally* valuable, can identify a much larger range of considerations that are non-fundamentally but nevertheless *noninstrumentally* valuable – that is, considerations that do not derive their value from being a means to some other end. Furthermore, there is considerable overlap among the considerations identified as noninstrumentally valuable by different normative theories, even if the fundamentally valuable considerations are radically different. Because, according to Markovits, agents are praiseworthy when the reasons that motivate them are the same as the reasons that their actions are right – and because all noninstrumental reasons are included as right-making – this allows for the possibility that agents can be praiseworthy even when they follow the wrong normative theory.[46]

In offering these suggestions, both Hurka and Markovits are apparently primarily interested in finding a way to attribute virtue and/or praiseworthiness to agents who follow the wrong normative theory. This is not a motivation for my account. Even so, at least one of the problem cases I describe, Torture, seems to be similar to those that interest Hurka and Markovits. And their proposed schemas could reasonably be understood as parallel attempts at resolving this kind of problem case; both propose a sort of multi-level structure[47], and seem to be making a claim about which levels are relevant to virtue and praiseworthiness.

As I hope the previous chapter made clear, the problem cases of interest here are not limited to those that involve false beliefs about normative theory. They include the case of Newman, who has no attitudes at all towards moral abstracta, as well as that of George, who has the correct attitudes towards the higher-level features but not towards the lower-level ones. And

46 pp.228-229.
47 Although only Hurka refers explicitly to multiple levels.

the development of a method for solving these cases requires a much more extended discussion than is offered either by Hurka or by Markovits; it is, after all, the main task of this dissertation. Even so, their remarks should be acknowledged as prefiguring, at least partially, the project being undertaken here.

III. Right-Making Features and Normative Explanation

I hope to have established that there are sometimes multiple features that make a particular action right or wrong. One might worry, however, that I have not succeeded in establishing an important, further claim – the claim that there is no basis on which to identify any of these features as the "real" right- or wrong-making features at the expense of the others. This claim is important, because, if it were false, then we would not require a further, substantive story about what counts as a desire for the actual good – we could simply identify desires for the actual good as desires for the "real" right-making features. In this section, I address two possible ways of identifying some right-making features as "real" at the expense of the others; I ultimately argue that both strategies are unsuccessful.

Because it is the simplest of the problem cases, I use Torture for the purposes of illustration in this section. By way of review, here are the three features which I identified as explaining why it would be wrong for Elaine to torture the prisoner:

(T1) Torturing the prisoner would constitute treating him as a mere means.

(T2) Torturing the prisoner would cause him pain.

(T3) Torturing the prisoner would cause certain events to occur in his brain.

My claim is that T1-T3 all explain why the action would be wrong, and thus all have equal standing as wrong-making features. But one might object in one of two ways. First, one might claim that *explanation* is not sufficient for a feature to be wrong-making in the relevant

sense. Instead, one might claim, we should look for a feature that *grounds* the wrongness of the action; and, since we might argue, there can only be one such feature, there will be only one feature that is wrong-making in the relevant sense. Second, even if we do not require that a feature *ground* the wrongness of an action in order to be wrong-making, we might still require that it do explanatory work of a distinctively normative kind. And, if only one feature of each action does the relevant kind of normative explanatory work, then each action will have at most one genuine wrong-making feature.

I begin with the appeal to grounding. The precise details of the grounding relationship are subject to debate, and most need not concern us here.[48] For our purposes, grounding can be understood as a one-way explanatory or "in-virtue-of" relationship between two relata which consist of facts or sets of facts. Significantly, one relatum does not ground the other unless the obtaining of the facts in the first relatum are *sufficient* for the obtaining of the facts in the second. So, for example, the fact that P does not itself ground the fact that P and Q. Instead, the fact that P and Q is grounded jointly by the conjunction of the fact that P *and* the fact that Q.

Regarding the features of actions, one might make the following claim. In order for a feature to be genuinely right- or wrong-making, the fact that that feature is present must *ground* the fact that the action is right or wrong. This will have the effect of eliminating most of the putative right- or wrong-making features that I have identified, as the presence of most of these features does not in itself ground the deontic status of the action. While features like T3 and T2 might *explain* the wrongness of torturing the prisoner, at least in some sense, they do not *ground* its wrongness because they are not in themselves sufficient to make it wrong. The fact that torture causes the prisoner pain makes the action wrong; but only in *conjunction with* the further

48 See, e,g Schaffer (2009), Rosen (2010) and Fine (2012) for general discussions of the grounding relation.

fact that inflicting pain constitutes treating someone as a mere means. The fact that torture causes certain brain states to occur makes the action wrong; but only *in conjunction with* the fact that those brain states cause or are constitutive of painful experiences. Only one feature, we might argue, is *sufficient* to explain the wrongness of the action – the "highest-level" feature on the list, T1. Thus we might claim that only the presence of this feature – the fact that the action involves treating the prisoner as a mere means – grounds the wrongness of the action, and that only this feature is genuinely wrong-making.

The problem, as Pekka Väyrynen (2013) discusses at some length, is that features like T1 apparently do *not* ground the moral status of actions. For T1 is not sufficient to make the action wrong. It requires the truth of a further fact – the fact that treating others as a mere means is wrong. In universes that are non-Kantian, actions that treat others as a mere means might not be wrong – this feature would not, for example, be wrong-making in a utilitarian universe. So, while it may seem promising, the appeal to grounding does not seem as though it will allow us to identify a single feature as right- or wrong-making. If a right- or wrong-making feature must genuinely *ground* the moral status of an action, then we will be forced to conclude that there are *zero* genuine right- or wrong-making features – for it seems that there are no features of actions that are *sufficient* to make them right or wrong without the truth of an additional fact.[49]

Two caveats are important here. First, Väyrynen introduces an important distinction between *bearers* and *sources* of normativity; and my analysis is premised on the view that the right- and wrong-making features being discussed here are bearers rather than sources. The idea

49 We might instead appeal to *partial* grounding; one fact need not be *sufficient* to explain another fact in order to partially rather than completely ground it. The problem here is that it seems quite plausible that more than one of the features of interest partially grounds the deontic status of an action. We might try to identify a single feature as partially grounding, but this would be essentially equivalent to searching for the feature that performs distinctively normative explanatory work, a strategy which is discussed below.

is that bearers of normativity are the features in virtue of which actions have their normative properties, and sources are the explanations for why those in-virtue-of relations hold. No bearer of normativity could ground the deontic status of an action on its own – unless it was *also* a source of normativity. If a schema were developed on which an individual feature could be both a bearer and a source of normativity, the appeal to grounding might be worth reexamining; I mention this possibility merely for the sake of completeness.[50]

Second, although I have argued that the appeal to grounding is not likely to succeed – that we cannot identify the right-making features with those that ground rightness, as there are no such features – there is nevertheless something intuitively compelling about this suggestion. Although T1 may not properly *ground* the wrongness of torture, it does seem as though it could be different in a meaningful way from T2 and T3. It seems to connect the action more "directly" to its deontic status than do the the other features; and, although it does require the obtaining of an additional fact to explain why the action is wrong, it requires *less* additional facts than the other features.

An alternative strategy appeals to this general intuitive sense that some of the right-making features are different from others. I have argued that features T1 through T3 all *explain* why it would be wrong for Elaine to torture the prisoner. But, we might claim, the different features do explanatory work of different *kinds.* Significantly, it might turn out that only one of these features does explanatory work of a distinctively normative kind. If so, we would have some basis for claiming that *this* is the "real" right- or wrong-making feature.

Much of the discussion over different varieties of explanation has taken place in the context of the literature on grounding. As noted, it seems unlikely that right- or wrong-making

50 Leary (Forthcoming) may be developing a schema of this kind; I offer no evaluation of it here.

features genuinely ground the moral status of actions. Nevertheless, many of the observations made about varieties of grounding seem as though they can be applied to varieties of explanation more generically. Kit Fine (2012), for example, argues that there are multiple kinds of grounding which correspond to multiple kinds of explanatory work; significantly for our purposes, he distinguishes between *natural* and *normative* grounding; since we are interested in explanation more generally, rather than grounding specifically, for our purposes we can distinguish between natural explanations and normative explanations. And there is some intuitive justification for thinking that these represent two fundamentally different kinds of explanations. It does seem as though there are different kinds of explanatory "work" that a given feature might perform, and it may seem that the wrong-making features that I have enumerated play different kinds of roles in explaining why torture is wrong.

Some features, such as T3, do work that is apparently non-normative in character. The fact that an action causes certain neural states to occur explains the fact that the action causes pain. But it does no *normative* explanatory work. There is nothing *bad* about certain neural states, except insofar as these neural states cause or are constitutive of pain. In Torture, we might think, the sole feature which *normatively* grounds the wrongness of the action is T1, the fact that it treats the prisoner as a mere means. The badness of the action, for lack of a better expression, may seem to "reside in" the treating-as-a-mere-means. And to a first approximation, this seems to be the main requirement for a feature to perform distinctively normative explanatory work: Features that perform normative explanatory work do so because they "contain" the intrinsic goodness or badness that is reflected in the action's deontic status.

If only *one* feature of each action were wrong-making in the *normative* sense, then we would have a principled basis for treating this feature as the "real" wrong-making feature; we

would then have grounds for claiming that agents should be evaluated based on their responsiveness to this feature as opposed to the others. If this strategy could be generalized to all problem cases, and if we could always find a single feature which normatively explained the moral status of an action, then the ambiguity would disappear without any need for my proposed solution.

But I do not think that this appeal to normative explanation succeeds in providing an acceptable solution to the problem. Suppose that only those features that do normative explanatory work are genuine right- or wrong-making features. AG accounts will still be ambiguous, I contend, because there will often be *multiple* features which do normative explanatory work. In each of the problem cases discussed in this dissertation, I contend that there are at least two features that appear to be doing some normative work in determining the moral status of the action.

Return, for example, to Torture. As stipulated in this case, the correct normative theory states that actions are wrong when they constitute treating a person as a mere means. Also as stipulated, the nature of pain is such that inflicting it always constitutes treating a person as a mere means. As I am reluctant to delve too deeply into the details of any particular Kantian theory, I left this description somewhat vague, but there are a number of ways of filling in what it is about the nature of pain that makes this the case. Suppose, for instance, that all agents are rationally bound not to will themselves to be in pain; the painfulness of the action, then, would be the feature in virtue of which it frustrates the prisoner's self-directed rational preferences. In this case, it seems quite plausible that the action's painfulness is doing part of the *normative* work in making the action wrong. Pain is such that rational agents must always will themselves not to experience it; this seems to be practically equivalent to saying that pain is *bad*. This badness is a

vital part of the explanation for the action's wrongness, because it explains why agents cannot rationally will themselves to be in pain. Thus the painfulness of the action seems to play an explanatory role quite different from that of the fact that the action causes certain neural events; we have some reason to think that it contributes some distinctively *normative* explanatory work.

Return now to Psychopathy. Newman's action is ultimately wrong because it violates his victims' right not to be harmed. And one might think that this feature is the only one doing any normative work. After all, it is the violation of *rights* that is directly identified as wrong-making by the correct normative theory in this case, not the infliction of harm itself; furthermore, the *badness* of rights-violation is presumably an essential part of explaining the act's wrongness. But, on reflection, it seems that the violation of rights is not the *only* feature that does normative work. For there must be some explanation for *why* persons have some rights but not others – they have the right not to be harmed, for instance, but not the right not to be offended. And it seems that this explanation must be a *normative* one. There must be something about harm that makes it so that agents are entitled not to be harmed. And whatever this feature is, surely, will be doing some normative work – it will make it the case that harm is *bad* in a such a way that persons have a right not to be subjected to it.

Return, finally, to Clinic Bomber. The correct normative theory commands us to preserve the lives of persons, and the fact that an action saves persons clearly seems to *normatively* explain why that action is right. But there must also be some explanation for *why* certain things *are* persons and others are not; and, once again, this explanation seems as though it must be a normative one. Imagine that we are arguing with someone like George, who believes that genetic properties are sufficient to confer personhood. We are likely to point out that genetic properties are implausible as a basis for personhood, because there is simply nothing morally *important*

about having certain genes. If this rhetorical strategy seems reasonable, it suggests that there must be something morally important about the personhood-conferring properties themselves; this in turn suggests that the actual personhood-conferring properties, here stipulated to be psychological, must be doing some normative explanatory work. There seems to be something *good* or *valuable* about having feelings or an enduring sense of self, and this value must be part of what explains why saving persons is itself good.

I have argued here that we cannot eliminate the problematic ambiguity by privileging a single feature as the "real" right- or wrong-making feature. We cannot identify the real right- or wrong-making features as those that ground the moral status of actions, as none of them do this; nor can we identify them as the features that perform distinctively normative explanatory work, because multiple features do this. In the end, we will have to confront the fact that a given action can have multiple right- or wrong-making features, and that none of these features is more genuine than the others. Even so, I claim, there is a way for us to obtain unambiguous attributions of moral worth. In the next chapter, I explain how.

Chapter Four

The Lowest Level of Normative Explanation

I. A Taxonomy of Right- and Wrong-Making Features

I have argued that

> For each normative theory N, there is some subset S of the features that make actions right or wrong, such that an agent's character depends only on whether he has the appropriate attitudes towards the features in S, and such that the moral worth of an agent's actions depends only on whether or not they reflect the appropriate attitudes towards the features in S.

My goal in this chapter is to develop a procedure that will take a full description of the normative facts as an input and which will output a list of the features of actions that belong in S. The procedure should be *general*, rather than limited in scope to the problem cases described earlier; this means that it cannot focus on the details of the features in these cases, but must instead identify a certain *kind* of features as belonging in S. It will therefore first be useful to discuss *which* kinds of right-making features there *are*. I have claimed that we cannot identify any of these features as the "real" ones, at the expense of the others. But even so, the various features do seem to differ from one another in important ways, and it seems that we can group them meaningfully into distinct categories.

For reference, here again are the features that would make it wrong for Elaine to torture the prisoner:

(T1) Torturing the prisoner would constitute treating him as a mere means.

(T2) Torturing the prisoner would cause him pain.

(T3) Torturing the prisoner would cause certain events to occur in his brain.

One potentially significant difference between these features is their "distance" from the fact of the action's wrongness. Lower features (T2 and T3) require a greater number of intervening facts to make the action wrong. T1, in contrast, is relatively "close" to the wrongness of the action – all it requires to make the action wrong is a single additional fact, which is that treating people as a mere means is wrong. It might seem at first that this difference provides a useful basis for classifying the features – we can group them according to their "distance" or "proximity" to the deontic status of the action, or according to how many additional facts each requires to make the action wrong.

There is clearly something interesting about the fact that these features differ in their distance from the deontic status of the action. Unfortunately, this difference provides a poor basis for a formal taxonomy of right- and wrong-making features, because there may be cases in which the distance or proximity of a given feature varies depending on how we individuate the intervening facts. In Torture, for instance, I supposed that T2 made T1 the case, because the nature of pain is such that inflicting it always constitutes treating the victim as a mere means. On the articulation I provided, T2 requires this one additional fact in order to make T1 the case. It then requires one more additional fact – the fact that treating others as a mere means is wrong – to make the action wrong. So, given that it requires two additional facts to make the action wrong, let us say that it is *two facts away* from the moral status of the action. The problem is that we can analyze this case with a finer grain, and may find additional facts between T1 and T2. We may reasonably ask *why* it is that inflicting pain always involves treating the victim as a mere means. Here is a possible answer: The infliction of pain is something to which a person cannot rationally consent.[51]

51 See, e.g. Kerstein (2013) for an analysis of the Formula of Humanity that centers on rational consent.

This, however, would seem to introduce a new wrong-making feature:

(T1.5) Torturing the prisoner would treat him in a way to which he could not rationally consent.

If T1.5 is also a wrong-making feature, then T2 is apparently *three* facts away from the wrongness of the action rather than two. We need one fact to get from T2 to T1.5 (the fact that an agent cannot rationally consent to having pain inflicted), another fact to get from T1.5 to T1 (the fact that acting in a way to which a person cannot rationally consent constitutes treating him as a mere means), and then one more fact to get from T1 to the wrongness of the action (the fact that treating someone as a mere means is wrong). Whether or not we choose to include T1.5 in our breakdown of the wrong-making features seems as though it might be arbitrary; i.e. depending on the level of detail we choose to provide, we can offer a version that either includes T1.5 or that excludes it, without a substantive difference in the content of our analysis. This means that a given feature's distance from the moral status of an action will vary depending on how detailed our analysis is; this in turn seems to imply that a taxonomy based on distance from an action's moral status will be an unstable one.

We might alternatively try to categorize the right- and wrong-making features according to their *contents*. The specifics of content will of course vary widely across actions and across normative theories. It will here be useful to enumerate the right- and wrong-making features of the other problem cases, for the purposes of comparison. Recall the features of Torture:

(T1) Torturing the prisoner would constitute treating him as a mere means.

(T2) Torturing the prisoner would cause him pain.

(T3) Torturing the prisoner would cause certain events to occur in his brain.

Compare to a possible list of features that could explain why it is wrong for Newman to

perpetrate a fraud on his victims:

(P1) Perpetrating the fraud will violate the rights of Newman's victims.

(P2) Perpetrating the fraud will harm Newman's victims.

(P3) Perpetrating the fraud will cause certain neural states to be realized in the universe.

We can also offer a list of features that might make a particular act of saving (genuine) persons right in the universe of Clinic Bomber:

(CB1) The action would save persons.

(CB2) The action would save beings with psychological properties X, Y, and Z.

(CB3) The action would save beings with certain neural or functional properties.

A bit of care is required in this case, because these are *not* actually features of George's action; recall that George is *mistaken*, that he does not save any persons, and that his action is not right. These are, instead, the features of a genuine act of person-saving which *would* explain the rightness of that act; what we want to know is which of these features George would have to care about in order to deserve praise for *attempting* or *intending* to perform a right action.

With the morally-relevant features of these actions laid out side by side, we might propose the following: The features can be grouped into meaningful categories on the basis of the kinds of content they contain. Some have content that is “low-level” or “concrete”: T3, P3, and C3 all concern neurological events or functional properties. Others have content that is “high-level” or “abstract”. T1, for instance, concerns the abstract notion of treating someone as a mere means; P1 concerns the abstract notion of violating rights; and CB1 appeals to “personhood”, which is arguably a complex and fairly abstract property. Still other features lie between these two extremes, with content at an intermediate “level” and with a moderate degree of “concreteness”. T2, P2, and CB2 all concern the kinds of properties that agents are familiar

with before studying either moral theory or neuroscience. They concern, for example, the pain or harm caused by an action, or the fact that it saves a being with certain (presumably pretheoretically familiar) psychological properties.

We might therefore appeal to the degree of "abstractness" of right- or wrong-making features in order to characterize them. This categorization scheme improves upon the previous one – which grouped features based on their distance from or proximity to the deontic status of an action – because it employs an intrinsic property of the features themselves and therefore will not vary based on the grain with which we analyze an action. Even so, this scheme is unlikely to be adequate for our purposes. While the notion of different degrees of "abstraction" seems to be fairly intuitive, it is nevertheless difficult to characterize formally. And because there is a great deal at stake – recall that we ultimately want to use the account developed here to guide our real-world judgments of blameworthiness – I am reluctant to place too much weight on notions that cannot be formally characterized.

Fortunately, there is a third alternative that *will* be adequate for our purposes. For while we may not be able to define a group of features in terms of its distance from the deontic status of an action, we *can* confidently assert that certain features perform explanatory work at a low*er* or high*er* level than others. The reason for this is the one-way making-the-case relationship between the features; the presence of T2, for example, makes it the case that T1 is present, but not vice-versa. If the presence of feature X makes it the case that feature Y is present, but not vice-versa, then feature X performs explanatory work at a *lower level than* feature Y.

This relative lower-than relationship, coupled with the distinction between normative and non-normative explanatory work introduced in the previous chapter, will be sufficient to identify those features that belong in S. I will ultimately defend the following:

> **The Lowest-Level Normative Features View (LLN):** For a given action, the right- or wrong-making features in the relevant subset S are the ones that perform normative explanatory work at the lowest level.

In the previous chapter, I argued that we cannot privilege a single right- or wrong-making feature on the grounds that it alone does normative explanatory work – for, in the problem cases, there are multiple features that do normative work. It is nevertheless possible to distinguish between those features that do and do not perform normative work; it is furthermore possible to identify the lowest feature among those that perform normative work.

Some right- and wrong-making features perform explanatory work that is clearly non-normative in nature; these include T3, P3, and CB3. The fact that an action has certain effects on neural or functional states can play a role in explaining its rightness or wrongness, but it is not a *normative* role. There is nothing intrinsically morally significant about neural or functional states, except insofar as they realize or cause certain psychological states.

In each of the problem cases, I argued that there are at least two features that perform distinctively normative explanatory work. In Torture, for instance, both T1 and T2 perform normative work in explaining why torture is wrong, since each seems to "contain" intrinsic badness that explains the deontic status of the action. Since T2 explains T1, but not vice-versa, we can identify it as the feature that performs normative explanatory work at the *lowest* level. We can offer similar analyses of the other problem cases. In Psychopathy, both P2 and P1 seem to contain intrinsic badness that is reflected in the wrongness of the action, and so both seem to to do normative explanatory work; since P2 explains P1 but not vice-versa, it is the lowest feature performing such work. In the world of Clinic Bomber, both CB2 and CB1 seem to contain intrinsic goodness that would be reflected in a given instance of person-saving, and so both seem to perform normative explanatory work; again, because of the one-way explanatory relationship

between the two features, we can identify CB2 as the lowest one.

As noted, the exact number of features which perform normative explanatory work may vary depending on the fineness of the grain with which we examine them. My contention is that it is always the *lowest* of these normative explanatory features which is relevant, so my procedure does not require that there be exactly two such features in order to work – there could be more than two, or even a single feature (in which case it would automatically be the lowest).

For any given action and any set of normative facts, we should be able to determine which feature of the action performs normative explanatory work at the lowest level. We therefore have a way of classifying the right- and wrong-making features that should be generalizable across actions and normative theories. It remains to be shown that these features are the ones that are relevant to moral worth; this is the task to which I turn in the next section.

II. The Lowest Level of Normative Explanation

LLN entails that responsiveness to the features that perform the lowest level of normative explanatory work is all that is relevant to character and responsibility.[52] Agents who are improperly responsive to these features are thereby vicious, and blameworthy if their improper responsiveness is reflected in their actions; no other features are relevant either to moral character or to moral blameworthiness. Why think that these are the features that belong in S? I offer two arguments in support of this conclusion here. The first appeals to T.M. Scanlon's recent work on the connection between social relations and the attribution of blame. Scanlon's view is that to judge an agent blameworthy is to judge that his actions have impaired our potential to form social relations with him. If we take this intuitively-compelling view seriously, then the

52 From this point forward I often refer to the "lowest-level" features of actions; this should be understood to designate the features which perform normative work at the lowest possible level. (Note that I am *not* referring to features concerning such things as neural events and so on; although these are arguably the "lowest" of the right- and wrong-making features, they do not perform normative explanatory work and are hereafter ignored.)

lowest-level normative explanatory features seem to be the most likely candidates for inclusion in S; the display of indifference towards these features seems the most likely to compromise an agent's potential for relationships with others. The second argument appeals to the fact that attitudes towards the good *de dicto* are irrelevant to character and moral worth on AG accounts. This feature of AG accounts, I argue, reflects a general commitment to the irrelevance of attitudes towards "formal" moral features of actions; this commitment strongly suggests that attitudes towards all right-making features above the lowest-level normative ones should be excluded from S.

Blame and Relationships

I understand AG accounts to offer descriptions of *which* conditions an agent must satisfy in order to be blameworthy; my aim is to develop a *complete* AG account, and thus a complete description of these conditions. But this does not imply that such an account would answer every theoretical question about blameworthiness. AG accounts, as I understand them, describe the conditions under which agents are blameworthy for their actions while remaining neutral on precisely *what* blameworthiness *consists in*. They are therefore compatible with multiple accounts of what blameworthiness *is*. One recently influential account offers us assistance in determining which features belong in S.

T.M. Scanlon (2008) argues that

> to claim that an agent is *blameworthy* for an action is to claim that the action shows something about the agent's attitudes towards others that impairs the relations that others can have with him or her. T0 *blame* a person is to judge him or her to be blameworthy and to take your relationship with him or her to be modified in a way that this judgment of impaired relations holds to be appropriate.[53]

53 p.128.

This view is both intuitively plausible and, in my view, compatible with AG accounts of moral worth. Intuitively, it certainly seems that one of the interesting features of blame is the way in which it affects future interactions; it is not implausible to suppose that blaming simply *is* the judgment that such interactions will be impaired as the result of what an agent has done. And an AG account of the *conditions* for blameworthiness can easily accommodate a Scanlonian account of the *nature* of blameworthiness. AG accounts tell us that agents are blameworthy when and because their actions express the wrong attitudes towards the actual good and bad; Scanlon tells us that a judgment of blameworthiness is the judgment that an agent's action has expressed attitudes that impair his or her capacity for future relations with others. These two kinds of accounts could dovetail nicely, if we understand the attitudes that impair one's potential for future relationships to *be* inappropriate attitudes towards the actual good and bad.[54]

This is not the venue for a defense of Scanlon's account; for the sake of argument, assume that an account of this kind is correct. The truth of such an account has the potential to guide us in determining which features belong in S. Given the truth of an AG account, an agent is blameworthy just in case his action reflects the wrong attitudes towards the features in S; given the truth of a Scanlonian account, to be blameworthy is to have impaired one's potential for

54 In the first chapter, I ruled out Strawsonian accounts on which blameworthiness is parasitic on human blaming practices; we might worry that this would rule out Scanlonian accounts as well. However, I do not think that it does. We *could* interpret Scanlon's account such that human social practices determine which attitudes impair the potential for relationships – in which case the account would be rather Strawsonian – but we do not *need* to. We could instead understand there to be objective and society-independent facts about which kinds of attitudes actually impair the potential for relations; what I am suggesting here is that these could be identical to the inappropriate attitudes towards the actual good and bad identified as relevant by AG accounts. This interpretation of Scanlon's view will imply that agents or even societies can be *mistaken* about whether the potential for relations with certain agents is impaired; this may initially sound odd, but should seem more plausible after some consideration. Consider – the fact that a person is sexually active outside of marriage does not *really* impair our potential for relations with that person, although previous generations might have viewed it as doing so. Those who felt that their potential for relations with such agents were impaired were presumably demonstrating that there was something wrong with their *own* attitudes, rather than that something *about the agent in question* had impaired his or her capacity for future relations.

relationships with others by way of one's actions. To determine which features belong in S, we should therefore ask which right- and wrong-making features are such that to display the wrong attitudes towards them impairs one's potential for relationships. The most plausible answer, I wish to suggest, is that the features of interest are those that perform normative explanatory work at the lowest level.

This suggestion may be surprising. The higher-level normative explanatory features – which concern such things as persons and the violation of rights – seem morally important to us. And it may seem that an agent who displays no concern for such features is one with whom our potential for relationships would be badly compromised. We may feel that we would cease to trust a person if we learned that he did not care at all about violating rights; we may feel that we would be terrified of an individual who did not care at all about killing or saving persons. But I think that we should not put too much weight on these initial reactions. When we imagine someone who does not care about violating rights, we are most likely imagining someone who does not care about the *specific* rights that she violates, either – we are unlikely to imagine someone who is *merely* contemptuous of rights in the abstract, but rather to imagine some who fails to care about specific, *de re* rights, like the right not to be harmed, lied to, etc. And when we imagine someone who does not care at all about persons, we are likely to imagine someone who does not care about the *de re* personhood-conferring *properties*, either. It is therefore difficult to tell which kind of bad attitude is doing the work of compromising our potential for relationships with this person – is it her attitude towards the higher-level feature, or the lower-level one?

We can gain some insight by considering some rather extreme cases in which agents respond properly to the higher-level features but not the lower-level ones. For instance:

> **Nazi Theoretician**: A ranking member of the Nazi Party is convinced that Jews are not people, and develops an elaborate theoretical explanation for this: Certain genetic properties, which he believes are possessed only by non-Jews, are necessary for personhood. He is deeply concerned with preserving those beings which he believes to be persons, but, of course, he is significantly mistaken about which properties make a being a person. Late in the war, he is assigned to oversee a concentration camp where he has a number of Jews killed; he does not feel bad, because he does not believe that they are persons.

Formally, Nazi Theoretician is quite similar to Clinic Bomber; both represent cases in which an agent is deeply concerned about persons yet mistaken about which properties confer personhood. The practical difference is that our intuitions about Nazi Theoretician are likely to be much clearer. I take it that all of us would consider our potential for future relationships with the Nazi described here to be radically compromised. This is significant, because the Nazi does *not* show any inappropriate attitudes towards personhood *qua* personhood; what compromises our potential for future relationships must be his attitudes, or lack thereof, about the personhood-conferring properties – the Nazi does *not* care about the properties that *actually* confer personhood.

Another example:

> **World Controller**: As a World Controller, Mustapha Mond is responsible for seeing that the rights of the millions of people under his jurisdiction are respected, a responsibility which he takes very seriously. He works hard to make sure that each citizen has a well-defined social role prepared for him or her, that everyone has access to soma, and above all that no one is exposed to ideas which might be frightening or upsetting. But Mond is deeply mistaken about which rights the people under his charge actually possess. In fact, they possess a right to autonomy, and Mond's actions ensure that this right is systematically violated.[55]

Mond cares deeply about rights *qua* rights; he has the right attitude towards a higher-level right-making feature of actions. But he has the wrong attitude towards the individual, actual rights to which his citizens are entitled. Our reaction to this case may be weaker than our reaction

55 This example is inspired by Aldous Huxley's (1932) *Brave New World.*

to Nazi Theoretician, but I take it that most of us would still consider our potential for relationships with someone like Mond to be severely impaired.

It will not be possible to survey every possible case here. But in general, it seems that we are more likely to view inappropriate attitudes towards the lower-level features, rather than those towards the higher-level features, as relationship-impairing. If a Scanlonian account of the nature of blameworthiness is correct, therefore, we have some reason to believe that it is the lower-level features, but not the higher-level ones, that are contained in S and therefore relevant to moral worth.

Moral Virtue as Moral Competence

The second argument does not require us to assume the truth of a Scanlonian account; instead, it appeals to a major motivating assumption that underlies AG accounts themselves. Recall that AG accounts are designed to exclude attitudes towards the good and bad *de dicto* as irrelevant to moral worth; this is what allows them to produce plausible results for Huck Finn and in other similar cases. I argued in the first chapter that the ability of AG accounts to handle these cases in an intuitively plausible way was a major consideration in their favor. My contention is that the same reasoning that leads us to exclude attitudes towards the good *de dicto* also commits us to excluding attitudes towards right- and wrong-making features other than those that perform the lowest level of normative work.

To show why, it will first be helpful to comment on an aspect of AG accounts that I have not previously discussed. On such accounts, moral virtue can be understood as representing a particular kind of competence. Let *competence with respect to domain X* consist in X-appropriate responsiveness to the considerations that an agent encounters while acting in his capacity as an X-agent. The idea of competence at work here is meant to be general, and there are a range of

domains that can stand in for “X”. Arpaly compares moral competence to artistic and business competence[56], but I consider a medical analogy to be clearer. Agents who are doctors are medically responsible for responding to certain considerations in a medically appropriate way. When a patient presents with certain symptoms, a good doctor will order the interventions that are appropriate to those symptoms; which interventions are appropriate is, presumably, determined by which ones will have the best effect on the health of a patient. A doctor's proclivity to respond in a medically appropriate way to the features of her patients represents her *medical competence*; her medical competence can be understood to be a measure of her quality *qua* doctor. A doctor who failed to respond in a medically appropriate way to the symptoms of her patients would thereby be demonstrating a defect in her quality as a doctor, and would also be an apt target for distinctively medical sanctions (liability to malpractice suits, the suspension of her professional license, etc.).

Moral virtue can be understood as competence in the moral domain. Certain considerations give agents moral reasons to act in certain ways; the quality of an agent *qua* moral agent depends on whether he is appropriately responsive to the considerations that give him moral reasons to act. When he is not, he demonstrates that he is a morally defective or vicious agent, and, if this is reflected in his actions, he makes himself an apt target for distinctively *moral* sanctions – blame and punishment. Levy (2007) has notably criticized the idea that moral virtue is a kind of competence. He points out that we generally do not blame those agents who display incompetence in non-moral domains – we would not *blame* a bad artist for failing to respond to her artistic reasons, for instance. But Levy has misunderstood the analogy between moral competence and competence in other domains. The analogy does *not* imply that those who

56 2003, pp.172-3; 2006, pp.34.

demonstrate incompetence in non-moral domains should be subject to *moral* blame, as moral blame is the sanction appropriate for distinctively *moral* failures. Instead, it implies that those who demonstrate incompetence in non-moral domains are apt targets for whatever blame-like sanctions are appropriate to those domains. These include distinctively medical sanctions, such as liability to malpractice suits, in the case of medical incompetence; in cases of legal incompetence, they might include disbarment; in the case of artistic incompetence, they might include aesthetic criticism or even mockery.

It is important to note that desiring to do well *de dicto* in various domains of human endeavor generally does not constitute being competent in those domains. It may contribute *causally* – desiring to be a good artist or a good doctor can cause one to work to develop the relevant competencies – but a person is not a good doctor or a good artist *in virtue of* desiring to be one. To be competent in one of these domains requires the appropriate attitudes towards the *specific considerations* that are relevant in that domain. Suppose that the medically appropriate response to a patient who presents with abdominal pain is to order a diagnostic X-ray. Part of being a good doctor is responding to these patients by ordering X-rays; and a doctor does not get any "credit", *qua* doctor, for *wanting* to perform the correct procedure *de dicto* without knowing what the correct procedure *is*. Morality, on AG accounts, is similar. One doesn't demonstrate any moral competence – and thus one doesn't demonstrate any virtue – by wanting to act well *de dicto*. To demonstrate moral competence, one must display appropriate responsiveness to the contents of morality – one must want to perform those actions which have the features that are actually right-making, and to refrain from performing those actions which have the features that are actually wrong-making.

So attitudes towards the good and bad *de dicto* are excluded because they do not seem to

be directed at the contents of morality and therefore do not seem to contribute or detract from an agent's moral competence. I argue here that attitudes towards higher-level right- and wrong-making features of actions – that is, attitudes towards any features other than those that perform the *lowest* level of normative work – should be similarly excluded, because they also fail to concern the contents of morality in the relevant sense. The notion of the "contents" of morality may seem rather vague, and we may reasonably ask what the "relevant sense" of such contents is. Here the analogy between moral competence and competence in other domains again provides a clue. Each domain has a set of considerations that provide reasons that are relevant to human endeavors in that domain. How do we determine which considerations and which reasons are relevant to a given domain? The answer seems to be that this is determined by the goal of the domain itself. The goal of medicine, presumably, is to make people healthy; the goal of art is to make works that are aesthetically good, and so on. And the considerations and reasons that are relevant to competence in a given domain appear to be those that are relevant to determining whether or how well an agent can accomplish the goal of a given domain. The reasons relevant to medicine are determined by which kinds of actions will promote the health of patients, and the considerations that are relevant are the considerations such that a doctor's responsiveness to them will make a difference as to whether he promotes health effectively; ditto for art, law, and other non-moral domains.

Significantly, however, not all considerations that are relevant to the *goal* of a given domain seem to be relevant to *competence* in that domain. Suppose, for instance, that an artwork is aesthetically good if it expresses the sublime; thus, the sublimity of an artwork is an aesthetically-good-making feature. And suppose that there are various features of artworks *in virtue of which* they are sublime. Since sublimity makes an artwork good, these sublime-making

features are, transitively, also aesthetically-good-making features. It does not seem that a would-be artist is a good artist simply in virtue of desiring to express the sublime, unless he also knows *how* to do so. The desire to express the sublime in one's works, without the knowledge of nor inclination to incorporate any of the sublime-making *features*, does not amount to artistic competence. Conversely, it seems that an artist can be a very good one without having any attitudes at all towards the sublime as such – so long as she cares about the sublime-making features and is motivated to incorporate them in her artworks.[57]

The reason for this, I propose, is that what counts as competence in a given domain is determined by which considerations in that domain are potentially *action-guiding*. One cannot simply decide to make sublime artwork and then do so; one can only accomplish this by deciding to incorporate certain features that in fact *make* an artwork sublime. Attitudes towards the higher-level features of good artworks – such as their sublimity – are not potentially action-guiding *in isolation* and thus do not count towards artistic competence. One's attitudes towards these features *is* potentially action-guiding in some contexts – part of writing *about* art well, for instance, may require one to recognize the importance of the sublime. But this is a distinct domain from the creation of art itself, and has a distinct, corresponding form of competence.

Why is this relevant to establishing LLN? The answer is that only the lowest-level features of actions are potentially action-guiding in the way required for responsiveness to them to represent moral competence. One can care about the higher-level right- or wrong-making features of actions, and be motivated by them. But these motivations will not translate into right

57 The example of artistic competence is a complicated one, since perhaps being a good artist requires technical skill in addition to concern for the good-making features of artworks. In this discussion, "artistic competence" should be understood to represent that *part* of being a good artist that consists in having the right attitudes towards the features of artworks that make them good. Arpaly and Schroeder (2014b) offer a similar aesthetic analogy, theirs concerning literary taste.

actions *unless* one is also motivated to respond to the lowest-level features. One can care about persons, for example, and intend to protect them. But one cannot act on this intention unless one *also* has some account of what properties make a being a person, as well as corresponding attitudes towards those (possibly putative) personhood-conferring properties. Other higher-level features of actions will be similar. One's intention to respect or disrespect rights cannot be action-guiding unless coupled with an intention to respect or disrespect some *specific* right (or putative right), such as the right not to be harmed. Nor can one's intention to treat or refrain from treating someone as a mere means be action-guiding unless one has an account of which kinds of actions *constitute* treating someone as a mere means.

Note that this argument is not intended to apply merely to the high*est* level right- or wrong-making features. In the previous section I described how finer-grained analyses of particular actions could produce a longer list of right- or wrong-making features. A more detailed analysis of Torture, for instance, gives us T1.5 – the fact that the action treats an agent in a way to which he could not rationally consent. An agent's attitudes towards this intermediate feature also fail to be action-guiding, unless accompanied by attitudes towards the lowest-level feature – one cannot treat or refrain from treating someone in a way to which he could not rationally consent, unless one also treats or refrain from treating him in some *particular* way which would explain why this higher-level property would be present.

So it seems that only an agent's attitudes towards the lowest-level features which perform normative explanatory work can contribute towards his competence *qua* moral agent; thus, these are the only attitudes relevant to his character, and these features are the only ones that belong in S. It is important to note that this does not imply that an agent's attitudes towards the higher-level features are irrelevant for *all* purposes. Caring about these features might cause an agent to

further investigate the lower-level features and to form appropriate attitudes towards them, thus becoming morally more competent and more virtuous. And attitudes towards these features might be action guiding in *some* domains, even if not in moral decision-making. Attitudes towards the higher-level features might be action-guiding in moral *theorizing*, for instance – I assume that considering or defending a particular moral theory counts as a kind of action – and thus responsiveness to them might be part of being a competent moral *philosopher*, even if it is irrelevant to whether one is a competent moral *agent*.

So we have two reasons to accept LLN. The first is that it is supported by a leading account of the nature of blameworthiness that is particularly consonant with AG accounts of the conditions for blameworthiness. The second is that the motivations of AG accounts themselves – with their commitment to the irrelevance of attitudes towards the good and bad *de dicto*, as well as to moral virtue as a kind of domain-specific competence – give us reason to think that only an agent's attitudes towards the lowest-level features are relevant to moral worth. In LLN, we have the procedure required to make an AG account complete, and to enable it to render judgments in the full range of previously problematic cases. In the next chapter, I describe a full, formal account of moral worth which incorporates LLN; I then demonstrate how it resolves the problem cases described earlier, and offer an overview of some of its practical implications. Before moving on, however, it will helpful to address one outstanding worry about LLN – might this account identify too many features as belonging in S?

III. Conceptualization and LLN

For each of the problem cases, I claimed that there were at least three genuinely distinct right- or wrong-making features. But one might worry that even if *some* of these features are genuinely distinct from one another, others are not. Specifically, one might worry that each of the

lowest-level normative features – features like T2, P2, and CB2 – is identical to some even lower feature. T2, for instance, concerns the pain that would be caused by torturing the prisoner. On some accounts of mind, psychological properties are identical to neural properties; if such an account of mind is correct, then T2 will turn out to be identical to T3, which concerns the neural events that the action would cause.

This poses a problem, because I have claimed that the features in S, which are relevant to moral worth, are those which perform normative work at the lowest possible level. As I will argue in the next chapter, the features which perform such work, and therefore belong in S, are features like T2, P2, and CB2; an agent must display an attitude towards these features in order to count as displaying an attitude towards the actual good or bad. But if T3, P3, and CB3 are identical to these features, then it seems that they must belong in S as well. If the fact that an action causes pain performs normative work at the lowest level, and the fact that an action causes certain neural states is the same fact as the fact that it causes pain, then the fact that the action causes certain neural states must apparently also perform normative work at the lowest level. The inclusion of these features in S would mean that attitudes towards them can make a difference to moral worth; yet, as noted earlier, it seems clear that attitudes towards neural and functional states cannot count as virtuous or vicious, at least in isolation.

It is important to note that this worry depends on the claim that features like T3, P3, and CB3 really *are* identical to higher-level features, rather than distinct features themselves. And this is a claim which we are by no means compelled to accept, for the relationship between neural and psychological events need not be one of identity. On non-physicalist theories of mind, neural events can be understood as *causing* psychological ones, for instance.[58] Even on

58 See, e.g. Chalmers 1996.

physicalist theories, there may be good reason to view neural and psychological events as distinct; this seems especially true on functionalist accounts, which allow that a particular psychological event can be realized by a variety of physical systems. My own inclination is towards some account of mind on which neural and psychological events and properties are *not* identical, so the worry described here does seem particularly worrisome from my perspective.

But suppose that we think that neural properties and psychological properties *are* identical, and thus that some of the lowest-level features which do normative work are identical to apparently irrelevant features concerning neural events. There is still a way to avoid the implication that our attitudes towards these neural events are relevant to moral worth, though it will require a modification to LLN. Specifically, we might adopt something like the following:

> **The Augmented Lowest-Level Normative Features View (ALLN):** For a given action, the attitudes relevant to moral worth are those which are directed towards the right- or wrong-making features that perform normative explanatory work at the lowest level and which are correctly conceptualized.

ALLN can be understood as a kind of "hybrid" account, which requires us to ask *both* which features perform normative work at the lowest level *and* which conceptualization of those features is relevant to moral worth. If we are worried that our account will identify too many attitudes as relevant to moral worth, ALLN should provide an adequate solution – attitudes towards neural states and events would presumably be excluded because they target the right- or wrong-making features under the wrong conceptualization. And ALLN still allows us to make unambiguous judgments of moral worth in the problem cases, because attitudes towards the features which perform normative work at higher levels are also excluded.

The switch to ALLN would come with one significant cost, however, because it would compel us to develop some account of the "correct" conceptualization of the right- and wrong-

making features. We may not require a fully-worked out account for many practical purposes – whatever the correct conceptualization is, a conceptualization of pain as a certain neural state is clearly an *in*correct one, and the ability to exclude such clearly irrelevant conceptualizations may allow us to evaluate most actions successfully. Even so, we might desire a complete account of which conceptualizations are correct, both for theoretical purposes and because we may worry that some conceivable problem cases could turn on more difficult questions of conceptualization. We cannot look to the original version of LLN for guidance, as it does not invoke conceptualization and thus does not incorporate any account thereof. Nor, for the reasons offered in Chapter Two, should we rely on Arpaly and Schroeder's procedure, according to which the correct conceptualization is identified directly by the correct normative theory.

However, the discussion in the previous section may once again be of assistance. My task there was to determine which right- and wrong-making features are relevant to moral worth. To this end, I identified two desiderata that the relevant right- and wrong-making features should satisfy: They should be such that incorrect responsiveness to them seems to compromise an agent's potential for relations with others, and they should be such that correct responsiveness to them represents moral competence. Our present task is to determine which *conceptualizations* of the relevant features are relevant to moral worth, but the same desiderata seem likely to be of use to us. An agent who responds incorrectly to the *relevant* (or "correct") conceptualizations of the right- and wrong-making features should be one whose potential for relations with others seems compromised. And an agent who responds correctly to the relevant conceptualizations should seem to be one who is morally competent.

I will not give a full accounting here of precisely which kinds of conceptualizations are identified by these two desiderata. To do so would require a full taxonomy of which kinds of

conceptualizations there *are*, and, as with the previous taxonomy of right- and wrong-making features, this would represent a lengthy and substantial discussion of its own. But I presume that these desiderata *could* identify a unique set of conceptualizations as correct. In any case, it should be clear that they allow us to exclude attitudes towards neural states as *incorrect*. An agent's attitudes towards neural states are not apt to compromise his potential for relations with others, unless he knows which psychological states they correspond to; nor are an agent's attitudes towards neural states apt to be action-guiding in the way required for his responsiveness to them to represent moral competence.

To reiterate, the move to ALLN is only necessary if we believe that the neural features of actions are identical to certain of their other right- or wrong-making features. Since I believe that these features are best regarded as distinct, I retain the original version of LLN in the remainder of this dissertation. Readers who prefer ALLN, however, are free to substitute it when necessary – all future claims about attitudes towards the lowest-level normative features can be understood as claims about attitudes towards the lowest-level normative features *under the correct conceptualization*.

Before moving on, I pause to address an additional outstanding question. The worry addressed in this section was motivated by the possibility that *some* of the putatively distinct right- or wrong-making features might turn out to be different conceptualizations of the same feature. But what if it turns out that *all* of the putatively distinct features are different conceptualizations of the same feature? What if, for instance, CB1, CB2, and CB3 all are identical to one another? I claimed in Chapter Two that this is not likely to be the case, but also noted that readers who preferred to treat these as different conceptualizations of the same feature were free to do so. I note here that ALLN should work even if it turns out that *all* of the features

in each case are identical. If CB1, CB2, and CB3 are all different conceptualizations of the same feature – call it CB* – then it follows that this single feature is the one which performs normative work at the lowest level. We must then determine which conceptualization of CB* is the one relevant to moral worth. As noted, it seems that the desiderata described in the previous chapter will allow us to make this determination. We must ask which conceptualization is such that faulty attitudes towards it compromise relationships, and correct attitudes towards it represent a kind of competence. These questions will point us to CB* under its conceptualization as CB2, and the end result will be the same as that provided by LLN.

Chapter Five

A Complete Account of Moral Worth, and an Overview of its Implications

I. Putting It All Together

In the preceding chapters, I developed a procedure for determining which attitudes count, for the purposes of assessing moral worth, as attitudes towards the actual good or the actual bad. To do any work in assessing agents or their actions, however, this procedure will need to be fitted into a full AG account. This is the first goal of the present chapter. The second goal is to demonstrate that this strengthened account can handle the cases which were previously problematic; the third is to provide an overview of some of this account's implications.

I propose the following account of moral worth:

MW: Agents are morally *praiseworthy* for free actions that reflect one or more of the following:

a.) concern for, a motivation to promote, or an otherwise appropriate pro-attitude towards the features of actions that make them right or good and which perform the lowest-level of normative explanatory work;

b.) abhorrence for, a motivation to discourage, or an otherwise appropriate anti-attitude towards the features of actions that make them wrong or bad and which perform the lowest level of normative explanatory work; or

c.) a lack of concern for, a lack of motivation to promote, or a lack of other inappropriate pro-attitudes towards the features of actions that make them wrong or bad and which perform the lowest level of normative explanatory work.

Agents are morally *blameworthy* for free actions that reflect one or more of the following:

a.) a lack of concern for, a lack of motivation to promote, or a lack of other appropriate pro-attitudes towards the features of actions that make them right or good and which perform the lowest level of normative explanatory work;

b.) contempt for, a motivation to discourage, or another inappropriate anti-attitude towards the features of actions that make them right or good and which perform the lowest level of normative explanatory work; or

c.) concern for, a motivation to promote, or an otherwise inappropriate pro-attitude towards the features of actions that make them wrong or bad and which perform the lowest-level of normative explanatory work.

It should be noted that I remain neutral as to precisely which pro- or anti-attitudes are relevant. In the preceding chapters I often adopted Arpaly and Schroeder's characterization of these attitudes as desires, but this detail is not important for our purposes, and I remain neutral in this formal account.[59] Given the exposition in the preceding chapters, MW should otherwise be more or less self-explanatory. The question to which I turn in the remainder of this chapter is whether and how we can put this account to work.

II. Solving the Problem Cases

The need for a strengthened account of moral worth was due to the fact that existing AG accounts were *incomplete,* and this incompleteness was demonstrated by the fact that they failed to produce unambiguous attributions of moral worth in several realistic problem cases. To test MW, then, we should apply it to these problem cases to see if it does a better job. For ease of reference, I reproduce each case here, along with the respective right- or wrong-making features of each.

> **Torture:** Elaine is a CIA agent, and is ordered by her superiors to torture a prisoner; she disobeys orders and refuses to do so, sacrificing her career. Elaine is a utilitarian, and believes that actions which fail to maximize utility are wrong; she concludes that torturing the prisoner would be wrong because it would cause more pain than it would prevent. Elaine's decision is motivated by three desires: the desire to act rightly, the desire to maximize utility, and the desire to avoid causing pain; she believes that all three desires can be satisfied simultaneously by refusing to torture the prisoner.
>
> As it turns out, torturing the prisoner *would* be wrong, but not for quite the reasons that Elaine thinks: The correct normative theory is a Kantian one, on which actions are wrong when and because they constitute treating an agent as a mere means. The nature of pain is such that to inflict it on a person always constitutes treating him as

59 It should be noted that I treat a *motivation* to act in certain ways as a variety of pro-attitude here.

a mere means. So the painfulness of torturing the prisoner does provide a reason not to do it, although this reason has nothing to do with utility.

Wrong-making features of torturing the prisoner:

(T1) Torturing the prisoner would constitute treating him as a mere means.

(T2) Torturing the prisoner would cause him pain.

(T3) Torturing the prisoner would cause certain events to occur in his brain.

Torture was problematic because it was unclear which attitude would need to be reflected in Elaine's action in order for her to be praiseworthy. Would praiseworthiness require a desire that the prisoner not be treated as a mere means, a desire which is *not* reflected in Elaine's action? Or would praiseworthiness merely require a desire not to inflict pain, a desire which *is* reflected in Elaine's action?

MW provides a straightforward procedure for evaluating Elaine. First we must list the right- or wrong-making features, which has already been done above. Next we must identify the features that do *normative* explanatory work; in this case, the two features of interest are T1 and T2. Then, we must identify which feature does normative explanatory work at the lowest level. This feature is T2; because T2 explains T1 but not vice-versa, T2 must be lower.[60] So Elaine's praiseworthiness depends on whether she displays an appropriate attitude towards T2. And, it seems, she does. Elaine wants to refrain from actions that cause pain to others; a desire to avoid wrong-making features is an anti-attitude of the appropriate kind. Thus, Elaine is praiseworthy.

60 To be clear: I assume here that T2 is the *lowest* feature which performs normative explanatory work. This seems like a reasonable assumption, as the *normative* explanation for why this action is bad seems to bottom-out in the fact that it is painful. Were there somehow an even *lower* feature which did normative explanatory work, then it would be this lowest feature that belonged in S. I make the same assumption about the two subsequent cases – i.e. I assume that P2 and CB2 are the low*est* level features which do normative work. If it could be shown that the normative explanations do *not* bottom out in these features, then my analyses of these cases would change.

Psychopathy: Newman is a psychopath. His general intelligence is higher than average, and he has a particularly good understanding of the psychology of others, which allows him to manipulate people very effectively. Newman has just perpetrated a financial scam, accepting large "investments" under false pretenses and then absconding with the money. Newman's action was motivated solely by a desire to enrich himself; he was aware that doing so would cause harm to others.

Although Newman's childhood therapist lectured him on why it is wrong to harm others, he did not (and does not) understand how the harmfulness of an action could provide reasons for him; nor does he understand that other people have rights, or why these rights should factor into his own deliberations. As it turns out, Newman's action *is* wrong. Individuals have a number of rights, including the right not to be harmed; the correct normative theory states that any action which violates the rights of others is wrong.

Wrong-making features of perpetrating the fraud:

(P1) Perpetrating the fraud will violate the rights of Newman's victims.

(P2) Perpetrating the fraud will harm Newman's victims.

(P3) Perpetrating the fraud will cause certain neural states to be realized in the universe.

Newman cannot understand what rights are, nor the fact that others have them; thus, we are supposing, he cannot show *any* attitude, appropriate or inappropriate, towards the fact that his action violates rights. But he *does* know that he harms people, and does not care; thus he *is* displaying a lack of concern for *this* wrong-making feature. Which feature is relevant for assessing Newman's blameworthiness? The procedure provided above will work in this case as well. The two wrong-making features that do normative work here are P2 and P1. P2 explains P1, so P2 is the lowest; thus it is Newman's attitudes towards P2 that determine the moral worth of his action. And what attitudes does Newman's action reflect towards P2? The wrong ones: The fact that an action harms someone makes it *wrong*, and thus it is appropriate to have an anti-attitude towards such a feature; an agent should, e.g. abhor inflicting harm. But Newman is indifferent towards whether he causes harm. Thus, his action displays indifference towards a

wrong-making feature of the relevant kind, and MW unambiguously judges him to be blameworthy.

> **Clinic Bomber:** George is deeply concerned with the well-being of persons, and strongly desires to prevent persons from being killed. Because he believes that fetuses are persons, he believes that he can save persons by preventing abortions. Accordingly, he places a small bomb in an abortion clinic and detonates it at a time when he knows the clinic will be unoccupied. The resulting damage to the facility, which is located in an area with limited access to abortion, forces it to close for several weeks and prevents a number of abortions which would otherwise have taken place.
>
> George originally became convinced that fetuses were persons when he read a description of their biology. A fetus has a complete and unique human genome, and this, George thinks, endows it with personhood. This belief is false – a being actually requires certain psychological properties in order to be a person – but George came to acquire it by reasoning responsibly and without self-deception. Fetuses do not in fact have the required psychological properties, so George's action does not actually save any persons. Even so, saving persons *is* morally good, and, if fetuses *were* persons, George's action would have been morally right.

The right-making features of a genuine act of person-saving:

> (CB1) The action would save persons.
>
> (CB2) The action would save beings with psychological properties X, Y, and Z.
>
> (CB3) The action would save beings with certain neural or functional properties.

The question here is whether George is praiseworthy. On the one hand, his action reflects a desire to save persons, which is actually morally good; on the other hand, his action reflects no attitudes at all towards the features that *make* persons persons. Whether George is praiseworthy will depend on which of these things a praiseworthy agent would need to care about. As before, MW provides us with an answer. Two features, both CB1 and CB2, do normative explanatory work. The lowest of these two features is CB2, and so we ought to evaluate George based on whether his action displays the correct attitude towards CB2. What is the correct attitude? CB2 is a right-making feature, so an agent should have some sort of pro-attitude towards it; one such

attitude might, for example, be a motivation to save beings with the listed psychological properties. George, as stipulated, does not believe that fetuses have these features. Therefore, his action does not reflect any attitudes towards these features, and he is not praiseworthy.

III. Beyond the Problem Cases

It seems that MW does provide unambiguous judgments of moral worth in the cases that were previously problematic. Of course, it is not possible to survey *all* possible cases to see if they are handled in a similarly unambiguous way. But I see no reason to think that the procedure described here is not generalizable, and I proceed under the assumption that MW is properly *complete* – given a normative theory and the attitudes reflected in an agent's action, it serves as a function that provides an unambiguous judgment of moral worth. MW has therefore accomplished what I set out to achieve in the introduction.[61]

Our *reason* for wanting a complete account of moral worth was not, however, merely theoretical. We wanted an account of this kind because judgments of moral worth are often both *difficult* to make confidently and of great practical *importance*. We may have to decide, for instance, whether a given criminal defendant deserves to be punished or should merely be quarantined from society; this question may turn on whether he is morally blameworthy.[62] It therefore seems rhetorically appropriate to conclude this project with a discussion of several of the practical implications of the account which I have developed. The remainder of this chapter is devoted to a discussion of some general implications of my view; the subsequent two chapters will address, in greater detail, two specific implications concerning psychopaths and

61 It is worth noting that MW should also *preserve* the desirable implications of existing AG accounts. It should, for instance, attribute praiseworthiness to Huckleberry Finn. The features to which Huck responds concern the harms that would be inflicted on his friend if he were to return to slavery, and these features presumably perform normative work at the lowest level.

62 Walter Sinnott-Armstrong has suggested, in personal communication, that our approach towards psychopaths should be one of quarantine rather than of punishment, on the grounds that they are not morally responsible.

ideologically-motivated wrongdoers, respectively.

The previous discussion of Torture illustrates one general implication of MW – it is in one sense rather *forgiving*, in that it often attributes praiseworthiness to agents even though they are mistaken about the correct normative theory. In this respect, MW does not differ significantly from existing AG accounts, which are already rather forgiving of agents who make explicit moral mistakes. In fact, as previously discussed, one of the major motivations for these views was to allow that agents like Huck Finn could be virtuous in spite of their false beliefs about morality *de dicto*. Arpaly and Schroeder (2014b) note that agents will often be praiseworthy in spite of their false beliefs about moral theory, because real-world agents often act as the result of non-theoretically motivated desires – such as a natural concern for the well-being of others – which are likely in many cases to track the actual good.

But there is another sense in which MW is *harsh* or *unforgiving*. For faulty responsiveness to the *actual good and bad*, on such views, is *constitutive* of bad character and thus cannot be excused. To a significant extent, this is also a feature of AG accounts more generally. But MW brings this feature of AG accounts into considerably sharper focus. The ambiguities that were problematic for previous AG accounts also obscured their implications; because it was unclear what exactly *counted* as appropriate responsiveness to the actual good or bad, it was difficult to confidently judge many agents as showing *in*appropriate responsiveness.

The most important general implication of MW is that many agents are morally much worse, and much more blameworthy for their actions, than we might originally have believed. One implication, for example, is that *psychopaths*, long the subject of controversy among philosophers, can be morally blameworthy for their bad actions. It is unclear whether there is a

prevailing view as to whether psychopaths are blameworthy[63], but there are at least a significant number of philosophers and psychologists who are strongly committed to the contention that psychopaths *cannot* be blameworthy. The entirety of the next chapter will be dedicated to a discussion of psychopaths, so I mention this topic only briefly here. We have already seen my analysis of one case of psychopathy, however – Newman's case – and it seems that this analysis can be generalized to other cases of psychopathic bad actions. Psychopaths are most likely unable to understand complex moral concepts concerning duties, rights, and reasons. But they most likely *can* understand the concepts required to express attitudes towards the features that do the *lowest-level* of normative work – features that concern harms, pains, property, etc. Because psychopaths are aware of these features, and do show inappropriate attitudes towards them, they are blameworthy according to MW.

This implication will be surprising to those who are committed to the view that psychopaths are *not* blameworthy. But perhaps more widely surprising is the implication that a large number of "normal" agents will be vicious – perhaps severely so – in virtue of their moral convictions, as well as correspondingly blameworthy for their resulting actions. For many active moral controversies, whichever party turns out to be wrong will be vicious – regardless of *which* side is wrong. And because of the extent of the disagreement surrounding many of these issues, we can safely conclude that *many* agents are wrong, and therefore vicious, without making any assumptions about *which* ones are wrong.

It will be helpful to take a step back for a moment. I have claimed that agents are vicious when they have inappropriate attitudes towards the right- and wrong-making features of actions

63 Interestingly, both sides of this debate have been known to claim that they themselves are in the minority. Maibom (2008, p.168), arguing that psychopaths *are* responsible, claims that "[m]ost philosophers" believe that they are not; Haji (2010, p.135), arguing that psychopaths are *not* responsible, claims that "[a] fairly dominant view" is that they *are*.

that do normative work at the lowest level; I have claimed that agents are blameworthy for those actions that reflect these inappropriate attitudes. In the remainder of this chapter, we are concerned with the question of *which* agents are vicious and blameworthy; we will therefore need to ask *which* agents have these inappropriate attitudes and act in ways which reflect them.

Some such agents are fairly obvious, and can be easily identified after making basic, commonsense assumptions about which considerations actually provide moral reasons. These are archetypally "bad" agents, who have attitudes which are pretheoretically recognizable as morally vicious. *Selfishness*, for example, is most likely an instance of inappropriate attitudes towards the relevant right- and wrong-making features. Presumably, the fact that an action will benefit someone else provides some moral reason to perform it; and, presumably, this feature is doing normative work at the lowest level. A selfish agent, who is unmotivated by the well-being of others, fails to have the appropriate concern for this feature and is therefore vicious; when this lack of concern is reflected in his actions, he is blameworthy. Sadism is another example: The painfulness of actions presumably makes them wrong, and presumably does so by performing normative work at the lowest level. A sadistic agent has *pro*-attitudes towards the infliction of pain when he should have *anti*-attitudes; thus he, too, is vicious and blameworthy for the actions that reflect this vice.

MW confirms our pretheoretical attributions of vice in these cases. But it also attributes vice and blameworthiness in a broader range of cases which may surprise us. Recall one of the general consequences of AG accounts described in the first chapter – while non-moral ignorance can excuse agents from blameworthiness, moral ignorance cannot. MW allows us to apply this principle more aggressively to real-world cases, as it allows us to identify more precisely what would count as an example of moral ignorance – for the purposes of assessing character and

blame, the kind of moral ignorance that cannot exculpate consists in ignorance of or false beliefs about which of the lowest-level features of actions are morally important. And when we examine many real-world cases of moral disagreement, we will find that many of the agents who are wrong are afflicted with moral rather than non-moral ignorance, and thus cannot be excused.

Here are two examples chosen because they are *realistic* (there are real-world agents who have the attitudes described), *dramatic* (the moral mistakes at issue seem to be potentially quite severe), and *important* (the moral matters involved are of great public significance).

(1) Some agents oppose abortion, and take actions to prevent or impede it, on the following grounds: Anything with a complete and unique human genome is a person, and we have a strong moral reason to protect persons. A fetus has a complete and unique human genome, and therefore we have a strong moral reason to protect fetuses.[64]

(2) Some agents oppose the use of contraception on the following grounds: It is morally bad to use a biological faculty for anything other than its intended purpose. Deliberately non-reproductive sex acts, such as those that employ contraception, use a biological faculty for other than its intended purpose; therefore, the use of contraception is morally bad.

To illustrate the general principle at work here, it will be useful to discuss each of these cases in some detail. To be clear: Unlike the vignettes discussed in previous chapters, these are not intended to describe thought experiments or hypotheticals. I assert here that there *are* agents who possess these beliefs and motivations, and furthermore that there are *enough* such agents for them to pose an interesting moral problem.

The putative moral significance of a complete human genome has not, to my knowledge, been cited in the philosophical literature on abortion. But it is important to realize that the public discourse on issues such as abortion is often divorced from philosophical discourse. And a survey of the public rhetoric on this subject demonstrates that the genetic properties of the fetus

64 This reasoning is essentially the same as that followed by George in Clinic Bomber.

are sometimes cited as a reason for wrongness of abortion. Consider, for example, the following passage from *Live Action News*, an online venue apparently dedicated to providing anti-abortion messages:

> Science teaches without reservation that life begins at conception. It is a scientific fact that an organism exists after conception that did not exist before conception. This new organism has its own DNA distinct from the mother and father, meaning that it is neither part of the mother nor part of the father... It is indisputably human, as it has human DNA... According to all the laws of nature, the unborn baby is human... Science declares that they are human beings with inherent value. The value of human beings is not dependent on where they are, how tall they are, what race they are, what they look like, or how old they are. Each person has inherent worth because of *who* and *what* he or she is: a member of the human species.[65]

The terminology used by this author is somewhat different, but the passage seems to reflect fairly clearly the same moral reasoning that is described in (1) – and, for that matter, in Clinic Bomber. A fetus, we are told, is a "human being"; the author seems to take it as analytic that human beings have "inherent value", so I take "human being" here to be equivalent to the thick and normatively laden concept of a "person." And the fetus is a human being *because* of its genetic properties; that these genetic properties are supposed to be doing normative work is made fairly clear by the author's repeated assertion that "science" tells us all we need to know about the moral status of the fetus. To remove any potential for confusion, the author makes it clear that the moral status of humans does *not* depend on anything *other than* their genetic properties.

This may strike us as a rather difficult view to defend, and some rather significant objections may occur to us immediately. For instance, the fact that a *distinct* genome is essential to personhood would seem to have strange and undesirable consequences for identical twins, who have the same genes. Would this view imply that twins are not persons? Or perhaps it would

65 Terzo (2013).

imply that one of each pair of twins is morally "disposable", such that we could kill one of the twins without compromising the existence of the distinct genome or the value of the person? I have no doubt that there are further problems with this view that would be discovered upon additional consideration.

It is important to note, however, that I do not take myself to be doing anything rhetorically illegitimate by citing a bad argument against abortion. The objective of this discussion is not to reach any first-order normative conclusions; my concern is with the *character* of agents and the *moral worth* of their actions. The existence of better or more reasonable arguments against abortion has no bearing on the fact that *some* agents endorse this rather bad one; and, of course, we want to know how to judge these agents.

Let us see what MW has to say about agents who disagree over abortion. We should consider two possibilities. First suppose that the agents described in (1) are *right* – it really is wrong to kill fetuses, because their genetic properties really do make them persons. Consider those agents who acknowledge that a fetus has a unique genome, but who do *not* believe that these genetic properties are important and do not form any anti-attitudes towards destroying the beings that possess them. On the view described in (1), the genetic properties of fetuses make it wrong to kill them; furthermore, these properties perform the lowest-level of normative work in explaining why it is wrong to kill fetuses. Therefore, the fact that an action kills a being with these genetic properties is a wrong-making feature in S. An agent who is indifferent to whether he kills beings with these genetic properties thereby shows indifference to one of the relevant wrong-making features, and is morally vicious. Suppose that such an agent takes actions to promote access to abortion, actions which reflect his lack of concern for genetic properties. If so,

the agent's vice is reflected in his actions, and he is blameworthy.[66]

The second possibility we should consider is that the agents described in (1) are *wrong* – genetic properties are irrelevant to personhood and to morality. At the very least, MW will tell us that the agents in (1) are *not virtuous* in virtue of their concern for fetuses, and that they would not be praiseworthy for any resulting actions. Preventing abortions saves beings with certain genetic properties, and the agents in (1) believe that this is a right-making property, but we are assuming that it is not. The actual right-making features at the lowest level presumably concern some other property or properties (as suggested in Clinic Protester, psychological properties seem to be plausible candidates), and an agent would need to be motivated by concern for *these* features in order to be praiseworthy for saving or for attempt to save persons.

So if it turns out that genetic properties are morally irrelevant, MW confidently tells us that the agents in (1) are not praiseworthy for preventing abortions. But are these agents *blameworthy*? A judgment of blameworthiness might seem plausible in light of the harm that these agents can sometimes cause. Their actions to prevent access to abortion might, for instance, harm women by restricting their autonomy, harm society as a whole by increasing the number of unwanted births, or even harm the safety and property rights of other parties if the agents resort to violent means. We might think that such agents would be morally blameworthy for causing any such harms. After all, they will have inflicted these harms in order to save beings with complete human genomes. Since human genomes are not morally important, they will have done harm without justification; and, in general, someone who does harm without justification seems to be blameworthy.

66 We are being asked to entertain a moral counterfactual here – the claim that genetic properties are what make human lives morally valuable – that may strike us as difficult to imagine. In the context of discussing psychopaths in the next chapter, I also discuss at some length the significant obstacles that we may face when reasoning about radical moral counterfactuals.

The story is a bit more complicated than this, however. On an AG account of moral worth, these agents must have displayed a morally bad *attitude* in order for them to be blameworthy. So far, all we have established is that their attitude – their desire to save beings with complete human genomes – is *not* morally *good*. Presumably it is a matter of moral indifference which genetic properties are instantiated in the universe, and it may seem that the desire to protect certain genomes is morally neutral. If so, it may be that these agents display neither good nor bad attitudes by acting and are therefore neither blameworthy nor praiseworthy for their actions. In the final chapter, I will argue that an agent who endorses and is motivated by the moral claim in (1) *is* vicious and *is* blameworthy for any resulting actions, assuming that he turns out to be wrong. But this argument is complex and requires one small but substantive extension to MW. As such, I postpone further discussion for the time being.

Turn now to (2). The moral claims described here are apparently endorsed by many agents. In fact, this view of contraception is an approximation of the official line taken by the Roman Catholic Church. Their stated position is that contraception is unnatural in some morally-charged sense and therefore morally wrong.[67] The unnaturalness of contraception cannot simply be a brute fact, and must be explained by some other feature of the action; my best understanding is that the unnaturalness is supposed to stem from the ostensible misuse of a biological function. The idea that biological functions are morally significant has been endorsed by other writers in the context of discussing homosexuality; Gerard Bradley and Robert George (1995), for instance, argue that homosexual acts are unnatural, and thus impermissible, because they use the reproductive system for an unintended end. As before, the question of whether these moral claims are plausible need not trouble us: our goal is to evaluate the agents who are convinced by

67 See Pope Paul VI's *Humanae Vitae* (1968), especially §10,13.

these kinds of arguments, rather than to engage with the first-order normative claims themselves.

Our analysis of this case may be easier if we map out explicitly the putative wrong-making features of contraception:

(C1): Contraception is unnatural.

(C2): Sex acts that use contraception employ a biological faculty for other than its intended purpose.

(C3): Sex acts that use contraception are non-procreative uses of a system that was selected by evolution for procreative purposes, or designed by God for procreative purposes, or otherwise has a history that explains why it is meant for procreation.

As with the other cases of interest in this dissertation, we can isolate at least three distinct (putative) wrong-making features here. C1 is presumably meant to be doing some normative explanatory work, as the agents in (2) will identify unnaturalness as intrinsically bad. C2 identifies the features in virtue of which an action is unnatural, and I assume here that these features are meant to be the ones performing normative work at the lowest-level – the badness of an unnatural act must ultimately originate from the features in virtue of which it is unnatural. And C3 picks out the features in virtue of which C2 might obtain; I take C3 not to be performing any normative explanatory work, so it is irrelevant for our purposes.

Suppose that the agents described in (2) are right – contraception really is wrong because it is unnatural, and it is unnatural because it involves using a biological faculty for other than its intended purpose. Suppose also that some other agents believe that C2 is a genuine feature of contraception – that is, they believe that contraception interferes with the function of a biological faculty – but do not believe that this is wrong-making, and are not motivated to refrain from using or promoting contraception. What would MW tell us about these agents? These agents would be vicious. They know of C2, which is a relevant wrong-making feature of using or

promoting contraception, yet they fail to form the appropriate anti-attitudes towards it. What if their lack of such anti-attitudes is reflected in their actions? MW then tells us that they are blameworthy.

Alternatively, suppose that the agents described in (2) are wrong – neither the "unnaturalness" of an action nor the fact that it uses a biological faculty in an unintended way are morally relevant. How should we evaluate the agents in (2) if they act so as to impede access to contraception? As was the case with (1), MW can confidently tell us that these agents are *not* virtuous or praiseworthy. The attitudes they express by acting are, at best, morally neutral. Whether such agents are *blameworthy* is, for now, an open and interesting question. There is, as with (1), some intuitive reason to think that these agents *are* blameworthy; assuming that it is harmful to prevent access to contraception, they will have done harm without justification. As before, however, the question of blameworthiness depends on whether vicious *attitudes* are reflected in an action. While I will ultimately argue that an action of this kind *does* reflect vicious attitudes and thus that the agent *is* blameworthy for it, this argument will have to wait until the final chapter.

IV. General Implications and Difficulties

There are many moral controversies about which seemingly well-meaning agents disagree. MW implies that in many such cases, the agents who turn out to be mistaken are *not* in fact well-meaning. Although they may desire the good *de dicto*, this is not constitutive of a good will. In some such cases, agents display indifference towards low-level features that are actually morally important; in so doing, they display a moral vice. In other cases, agents will be motivated by their moral concern for features which are actually morally neutral, rather than right- or wrong-making. These agents do not display a good will, either; the quality of will they

display is, at best, neutral.

The main point is this: If to be well-meaning is to possess a good will, then whoever turns out to be wrong about these controversies is *not* well-meaning. Whoever turns out to be wrong is either indifferent towards actual right- or wrong-making features, which is morally vicious, or concerned about morally irrelevant features, which is morally neutral at best. It is important to underscore that the failure of these agents to mean well need not be due to any epistemic irresponsibility or self-deception; they may hold their false moral beliefs sincerely and as the result of responsible reasoning. The conclusion that they fail to mean well is an implication of MW's account of what it is to have a good will. To mean well requires one to have the right attitudes towards the features in S, and, I have argued, the agents who turn out to be mistaken about these kinds of moral controversies will often fail to possess the right attitudes towards these features.

It is important to clarify that an agent can be wrong about a moral claim, and yet still mean well, if she is mistaken about *non*-moral facts. Some moral disagreements presumably do result from non-moral mistakes; in cases such as these, it is entirely possible that all parties really do mean well. One easy example: Many agents disagree about the moral desirability of practices such as hydrofracking. It is easy to imagine, however, that these agents are in agreement as to what is basically morally valuable; they simply disagree about the empirical question of what impact hydrofracking will have on these bearers of value. One party might believe that hydrofracking will significantly improve human happiness by stimulating the economy; the other might believe that its economic impact will be minor. One party might believe that hydrofracking will significantly reduce human happiness because of its effects on the environment; the other might believe that its environmental effects will be negligible. Of course, agents might be

blameworthy if they have come to have their false non-moral beliefs by way of epistemic irresponsibility, such as if considerations of personal gain were to incline them to form beliefs favorable to hydrofracking. The main point is that false beliefs about the permissibility of practices like hydrofracking are not themselves automatically indicative of a vice; viciousness would require a faulty response to the features that make hydrofracking right or wrong.

In principle, it should be easy to distinguish cases of non-moral mistakes (which can exculpate) from failures to have the right attitudes towards the features in S (which cannot). One complication, however, is that it is often difficult to tell what is going on with real-world agents. It is not always easy to determine which attitudes are reflected in another agent's actions, and, even if the agents gave us honest self-reports, it is not clear that human introspective access is good enough for these reports to be perfectly reliable. In any case, it is likely that real world agents will often have a mixture of attitudes, and that *some* will reflect moral mistakes and others non-moral mistakes. For example, an agent might have a false empirical belief to the effect that the environmental damage caused by hydrofracking will be relatively small. He might *also* be insufficiently motivated by environmental concerns, perhaps because he regards the environment as important only because of its effects on humans, and fails to afford it independent status as intrinsically valuable. If this agent goes on to support hydrofracking, his action will partly reflect a moral vice and partly reflect a (possibly) innocent empirical mistake. The blameworthiness of these agents will presumably be partially reduced in proportion to the extent that their actions reflect blameless empirical mistakes rather than faulty responsiveness to the right- or wrong-making features; while the details may still need to be worked out, the existence of these more complex cases does not pose a fundamental problem for MW. Nevertheless, it is important to bear in mind the possibility of mixed motivations when we consider actual agents.

Another possible problem is that, even when we are certain of an agent's attitudes, it may be difficult to tell which attitudes reflect non-moral mistakes and which ones reflect faulty moral responsiveness. The distinction is a clear one in theory, but some kinds of attitudes may be difficult to classify. A particular difficulty is posed by attitudes with *religious* content. Suppose that an agent is otherwise like George, but instead believes that he must save fetuses not because of their genetic properties, but because God has commanded him to do so. To assess this version of George, we will first have to ask whether the action *would* be right if God *had* commanded him to do it. Suppose that it would. If so, we would then have to determine what kind of explanatory work God's commands perform; for George to be praiseworthy, God's commands would have to do normative explanatory work at the lowest level. It is not obvious how we should make this determination. If God's word is the sole arbiter of morality, then it seems that God's command will be doing the *only* normative work and, *a fortiori*, the lowest level of normative work. Alternatively, perhaps there is a two-step process – some features of actions make them holy, for example, and then God endorses the holy actions, making them right. In this case, the lowest-level work would presumably be done by the features that make the action holy, and to be praiseworthy George would need to respond to these.

So, to evaluate these cases, it seems that we would first need to solve the central problem presented in the *Euthyphro*. If God does not in fact exist, we must also work through some rather difficult counterfactual questions: If God *did* exist, *would* His opinions of actions be among their right- and wrong-making features? This undertaking is beyond the scope of this dissertation. My intention was to provide a generalized procedure that would allow us to determine the moral worth of actions in all cases, and I take myself to have done so; this does not mean, however, that the procedure will be easy to *implement* in all cases.

There are many motivations that are not religious in nature and which seem relatively easy to analyze. And, for a significant number of moral controversies, it will turn out that many mistaken agents are not displaying the appropriate attitudes towards the features in S and will therefore not be praiseworthy. It is of course not possible to survey all or even a large number of these disagreements here. I have described two in some detail, and leave others open to future exploration. I will, however, end this chapter by listing a few additional controversies over which disagreements are likely to be due to faulty responsiveness to the right- or wrong-making features in S.

> **Capital Punishment:** It may turn out that sufficiently bad agents deserve to be executed, and that this provides a moral reason to execute them. On the other hand, it may not, in which case executing these offenders may be impermissible. For certain offenders, there will be an agreed-upon fact about their degree of guilt. Some agents will respond to this fact as though it gives a moral reason to execute the offender. Other agents will respond as though it does not. It seems that whoever turns out to be wrong will be responding incorrectly to a basic right- or wrong-making feature, and will thus be vicious and potentially blameworthy.
>
> **Homosexuality:** As noted earlier, it has been argued that homosexual acts are impermissible because they use a biological faculty for other than its intended purpose. Many agents will agree that the reproductive system is intended for procreation. Some agents will respond to this fact as though it provides a moral reason not to use the reproductive system non-procreatively; others will not.
>
> **Factory Farming:** Many agents will agree that many of the animals harvested to produce meat are kept in very poor conditions, and that these animals find their experience unpleasant. Some agents respond to this as though it provides a strong moral reason not to facilitate factory farming by eating meat. Others do not. This difference in behavior seems to result from a difference in responsiveness to a low-level feature of meat-eating – the fact that it contributes to the suffering of animals. If this does provide a moral reason not to eat meat, then non-vegetarians will have been responding inappropriately to a wrong-making feature; if it does not provide such a moral reason, then moral vegetarians will have been responding to a morally irrelevant feature.

One could go on. I hope to have shown that there are a significant number of moral disagreements that are due, at least in large part, to differences in responsiveness to the low-level

features, rather than to disagreement about which such features are present. And if an agent responds incorrectly to the low-level features of actions, he will not express a good will and will therefore not be well-meaning. In the final chapter, I return to offer a final missing piece in this story and to argue for a stronger conclusion. Agents who fail to respond to the actual right- and wrong-making features in S are vicious, and blameworthy when their vices are reflected in their actions. But how should we evaluate those agents who *respond* to morally irrelevant features *as though* they are right- or wrong-making? I will ultimately argue that moral responsiveness to non-moral considerations is also vicious, and that this vice too can ground blameworthiness. In the next and penultimate chapter, however, I pause to consider an agent who is morally responsive to nothing at all – the psychopath.

Chapter Six

Psychopaths and Imaginative Resistance

I. Introduction

As argued in the previous chapter, MW implies that psychopaths *can* be blameworthy for their actions.[68] Psychopaths may be cognitively limited, particularly with respect to their ability to understand the abstracta that do high-level normative work in making actions wrong. But they often have average or above-average general intelligence, and are typically socially sophisticated and capable of manipulating others effectively. This suggests that psychopaths do have a good understanding of the psychology of other agents; they can presumably understand that others have mental lives very much like their own, that they can feel pain and have their preferences frustrated, and that they can be harmed. These are the kinds of features that, on plausible normative theories, do the lowest-level normative work. Since psychopaths generally *do* know that these low-level wrong-making features are present, and are generally *not* deterred by them, they display inappropriate attitudes towards the features in S. This, according to MW, is sufficient for them to be blameworthy.

Many philosophers have argued that psychopaths are *not* blameworthy. If they are right, of course, it will turn out that MW has a false implication, and we will need to reject it. Levy (2007) explicitly invokes this implication as a reason to reject AG accounts; he considers it clear that psychopaths are not blameworthy, and argues that we have grounds to reject a theory that implies otherwise. My aim in this chapter is to show that we ought not to view MW's

68 The qualifier "can" is due to the fact that psychopaths need not *always* be blameworthy – they can be excused from blame for the same reasons as normal agents. A psychopath might, for instance, give someone arsenic under the false belief that it is sugar. I do *not* mean to suggest that psychopaths have a diminished level of responsibility, or that they are blameworthy in a narrower range of cases than normal agents.

implications about psychopaths as a *reductio*. Most of the arguments advanced against the blameworthiness of psychopaths have already been neutralized by the claims established in previous chapters. One major line of argument, for instance, holds that psychopaths cannot understand the moral dimensions of their actions in the way required for them to express ill-will; since the expression of ill-will is required for responsibility, psychopaths cannot be blameworthy. Arguments of this kind (e.g. Levy 2007, Shoemaker 2011, Nelkin 2015) generally focus on psychopaths' inability to understand *that* others are entitled to moral consideration; since they are unaware of the moral entitlements of others, they cannot show contempt nor any other attitude towards these entitlements when they act. I am willing to concede that psychopaths cannot understand that others are entitled to be treated in certain ways and thus cannot show any attitudes towards others *qua* moral patients. But, as I have argued, attitudes of *this* kind are not relevant to moral worth. The fact that an action harms a person or a moral patient, or violates someone's rights, is likely to be one of the higher-level wrong-making features. Moral responsibility merely requires that an agent show attitudes towards the *lowest*-level features that perform normative work, and psychopaths *can* show attitudes towards *these* features.

To defend against the argument from ill-will at greater length would simply be to reiterate the claims defended earlier in this dissertation; I therefore set it aside. There is, however, another argument that is worthy of an extended discussion. Variations of this argument are offered by Levy (2007) and Shoemaker (2011), and it is notable for several reasons. First, it does not rely on any premises that I have already rejected. Second, it is both ingeniously simple and apparently compelling, and it is potentially convincing even in the absence of any preconceptions about moral worth. Finally, the reason that the argument fails – and I do contend that it fails – is an interesting one, and understanding it will require us to engage more extensively with some

independently important questions concerning moral counterfactuals and the imagination.

Levy and Shoemaker propose a thought experiment in which normal human agents are supposed to be in an epistemic position analogous to that of psychopaths. The humans are informed – by extraterrestrial visitors with superior moral sensitivity – that it is wrong to step on grass. Humans do not understand and cannot be motivated by these grass-related moral reasons, due to their lack of sensitivity to them; we are invited to intuit that the humans in this case would not be blameworthy for stepping on the grass, and to conclude that psychopaths are not blameworthy either. But the appeal to this thought experiment fails, I argue, due to the effects of *imaginative resistance,* a phenomenon which interferes with our ability to imagine certain moral counterfactuals. When the grass case is understood in such a way as to be properly analogous to psychopathy, it incorporates a moral counterfactual of the kind that can be expected to provoke imaginative resistance. The best we will be able to do, I argue, is to imagine a non-consciously modified case that lacks the moral claim of interest; since this modified case will be disanalogous to cases of psychopathy, our intuitions in response to it do not support any conclusion about psychopaths.

My response to Levy and Shoemaker's argument will first require a somewhat extended discussion of imaginative resistance; the next section is devoted to this discussion. In the final section, I turn to Levy and Shoemaker's argument and argue that it fails; thus, we need not conclude that psychopaths are excused from blame, and my account of moral worth is protected from this objection.

II. Imaginative Resistance

The literature on imaginative resistance traditionally illustrates the phenomenon by way of short, fictional vignettes that are intended to evoke it. In keeping with this tradition, consider

the following story:

> **First Contact:** Scans of the planet indicated that its inhabitants had developed a complex society, with art, philosophy, and democratic institutions of self-government. This was all the more remarkable for the fact that these life-forms were single-celled, closely resembling enormous versions of the freshwater amoebas of Earth. Of course, these creatures were hideous, and the captain did the right thing when he ordered his crew to open fire, sterilizing the surface of the planet.

Most of us will readily entertain the non-moral elements of First Contact in an imaginative context – we will accept that there really are intelligent amoebas and interstellar spacecraft within the world of the story. In contrast, most of us will *resist* the moral claim that ugliness is a justification for genocide – we will we feel that we are *unable* to imagine this claim, or that the author of the story fails to *make it true* within her fictional world.

The causes of imaginative resistance, as well as the mechanisms by which it operates, are the subjects of ongoing debates.[69] I do not aim to resolve these debates here, but my arguments in the following section will require us to have a very general understanding of *what happens* when we resist a claim, as well as of *which cases* are likely to evoke resistance. With respect to the first question, Weatherson (2004) identifies several phenomena which may occur together when we experience resistance, two of which are particularly important to the following discussion. The first is an apparent effect on *what we imagine*. When we consider a problematic moral claim, we may be struck with the sense that, despite our best efforts, we ultimately fail to imagine its being true. For instance, we may feel that although we understand the moral claim in First Contact perfectly well, we do not really *imagine* it when we read the story. We may imagine some group of humans, or the narrator of the story, *believing* that killing ugly lifeforms is morally good. But we may find that the claim itself – that a killing really *is* good, precisely *because* it is the killing

69 See Walton (1994), Gendler (2000), Weatherson (2004), Stear (2015), Gendler and Liao (2016) for particularly useful discussions of the phenomenon and of the outstanding controversies surrounding it.

of an ugly being – eludes our attempts to imagine it, much as a square circle eludes our attempts at visualization.

The second phenomenon is an apparent effect on *truth within a fictional world*. The author of a work of fiction has wide-ranging authority to make claims true within the fictional world she describes, irrespective of their truth status in the actual world; within the world of the story, she can make it the case that the Confederacy won the Civil War, or that faster-than-light travel is commonplace. When we encounter claims that we resist, however, we are often struck with the sense that the author's power is *limited* – we feel that she *cannot* make certain claims true *even within the world of her story.* So, returning to the example of First Contact, it seems natural to accept that there really are intelligent amoebas in the world of the story, while insisting that it is *not* right to kill them, even within the fictional world.

These two phenomena are important to the following discussion, for they explain why imaginative resistance can be expected to have an effect on our moral intuitions about particular cases. If it *seems* that we fail to imagine the truth of a given claim, then it is likely that we *do* in fact fail to imagine the truth of that claim. And if a given claim *seems* false in the fictional world that we are imagining – even if the author asserts that it is true – then it is likely that we are imagining a world in which the claim really *is* false. The end result, I will argue, is that when we encounter imaginative resistance, we fail to imagine the case in question, instead imagining a different case which lacks the problematic moral claim; any intuitions we form will therefore be in response to this modified case, rather than the case as originally described.

I turn now to the second question: *Which kinds of cases* should we expect to resist? There are really two subsidiary questions here; one concerns the *contexts* in which resistance can occur, and the other concerns the *contents* of the claims that are likely to trigger it. With respect to

context, the existing literature has focused primarily on the emergence of resistance in response to literary fiction; it is essential to my argument that resistance is not limited to fiction but can also be encountered in response to philosophical thought experiments. At first glance, there seems to be little difference between philosophical thought experiments and the short, fictional vignettes that are typically used to illustrate resistance – both are of limited length and complexity, lacking well-developed stories and characters – and thus we have a *prima facie* reason to think that resistance *can* arise in response to thought experiments.

While these short vignettes strongly resemble thought experiments, one might nevertheless worry that we mentally engage with these two kinds of cases in fundamentally different ways. Gendler, for instance, distinguishes between the mental acts of *imagining* and *supposing*, and suggests that resistance may affect what we imagine but not what we suppose.[70] While we generally describe ourselves as *imagining* the claims we encounter in literary fiction, we are more likely to describe ourselves as *supposing* the truth of the claims we encounter in thought experiments. And perhaps there is something about imagination which uniquely suits it to produce resistance. Weatherson suggests that suppositions are generally "coarser" than imaginings – imagining that P requires us to fill in a variety of details about the world in which P obtains, while merely supposing that P does not.[71] Depending on the mechanics of how resistance is triggered, these extra details could explain why imagined vignettes evoke resistance but thought experiments do not.

But while the distinction between imaginings and suppositions seems to be a meaningful one, the kinds of thought experiments that are of interest here – those in which we are asked to

70 (2000), pp.80-81.
71 (2004), p.20, footnote 9.

form moral intuitions about the case described – are likely to require a mental activity that is more similar to imagining than to supposing. We can certainly use suppositions, rather than imaginings, for reasoning tasks that do not require us to engage with the contents of the propositions of interest; formal logic is the most obvious example. But when we consider a case for the purposes of forming moral intuitions about it, we must engage with the contents and we must seek to fill in, to a significant extent, the details of the world described. Later in this section, for instance, we will encounter a thought experiment in which jurors must decide whether to convict a woman who has killed her baby. To form an intuition about the correct course of action in a case like this, we must form a gestalt impression of the world in which the jurors and the woman are embedded, in order to intuit whether a conviction *seems* right or wrong in this scenario.[72] This, it seems, is quite similar to what we do when we imagine fictional vignettes, and thus the distinction between imagining and supposing does not give us a reason to think that thought experiments should be immune to resistance.

Nevertheless, a number of philosophers have pointed out that the *genre* of a given work seems to play an important role in determining whether it will evoke resistance.[73] Gendler (2000) characterizes the phenomenon itself in terms of the relationship between the reader of a work and the work's narrator; resistance, on this view, is when the reader chooses or is compelled to challenge the narrator's authority, treating the problematic moral claims as false beliefs of the narrator rather than as truths about the fictional world. On this understanding of resistance, it might be unclear whether the phenomenon can arise in response to thought experiments. There is no real "narrator" of a philosophical thought experiment, to whom a reader might attribute the

72 In the course of raising methodological concerns about the use of "outlandish" cases in moral philosophy, Elster (2011) similarly argues that we require extensive background information about the world of a thought experiment in order to form moral intuitions about it.

73 See, e.g. Brock (2012) and Liao et al. (2014).

moral claims that she resists. We might propose that the author of the thought experiment is analogous to its narrator, but this analogy is poor: While the narrator of a fictional case makes assertions about what is true in the fictional world, the author of a thought experiment is merely inviting us to imagine what would follow if a given set of claims *were* true in a fictional world. Does this difference preclude the possibility of resistance?

To address this worry, as well as to answer the question of which kinds of *contents* trigger resistance, it will be helpful to discuss the interaction of imaginative resistance with what Nils-Hennes Stear (2015) calls "qualifying contexts" – sets of additional facts or background conditions which, when added to an otherwise problematic vignette, cause our resistance to disappear. Here is a commonly-cited example of a case which is generally expected to evoke resistance, originally offered by Kendall Walton (1994):

> **Infanticide:** In killing her baby, Giselda did the right thing; after all, it was a girl.[74]

What is it about this case that we resist? Presumably, it is the implicit suggestion that the lives of girls are either morally valueless or less valuable than those of boys. It is notable that this moral claim is not explicitly stated; there is nothing in Infanticide which strictly implies this or any other claim about moral value. Even so, a natural reading of Infanticide results in our attributing this claim to the narrator. We all realize that there are, sadly, some people who really do believe that girls are less valuable than boys; in reading Infanticide, we assume that the narrator is one such person, and we regard his moral claim as false even within the fictional world being described.

It is important to note, however, that the addition of a qualifying context can eliminate

74 Walton (1994), p.37; title added.

our resistance. Consider Stear's example:

> **Patriarchy**: In killing her baby, Giselda did the right thing; after all, it was a girl. Since the Patriarchy Party had seized power, all girls faced horrific lives of state-sponsored sexual slavery. Giselda felt nauseous killing her child; doing what's right isn't always easy.[75]

Stear notes that we are not likely to experience resistance in response to Patriarchy. Why not? Clearly, the qualifying context – the extra information about the Patriarchy Party and the baby girl's likely fate – makes the difference. But how? What seems to happen is that we cease to interpret the narrator as believing anything objectionable about the moral value of girls; we instead interpret him as believing that it is better to kill someone than to allow him or her to lead a life of state-sponsored sexual slavery. We may or may not agree with this claim, but we are likely to find it much less objectionable than the claim that the lives of girls are intrinsically less valuable than those of boys, and are correspondingly less likely to resist it.

Infanticide and Patriarchy illustrate two important points about the kinds of claims that are likely to evoke resistance. First, the mere fact that a moral claim is false is not sufficient to evoke resistance. It may not, in fact, be true that anyone has a moral reason to kill her baby, but we can entertain a story in which some people do have such a reason without encountering resistance. What we resist about Infanticide, and what is absent from Patriarchy, is an implicit claim about what is *intrinsically* morally valuable, or about what ultimately grounds our moral reasons. We do not resist the claim that killing girls is morally good; we resist the claim that killing girls is morally good *because their lives are less valuable than those of boys*. Call those moral claims that do make assertions about intrinsic moral value or about the ultimate grounds of our moral reasons *basic* moral claims.

75 Reproduced from Stear (2015), p.3, originally titled “Giselda*”.

Second, the mere fact that we regard a basic moral claim as *false* is not sufficient for us to resist it; even those of us who do not agree with the moral claim implicit in Patriarchy will, I take it, be able to entertain the case without resistance. The moral claim in Infanticide seems in some sense "farther out", or less plausible, than the one in Patriarchy; in my terminology, we regard the claim in Infanticide as *radically* counterfactual. As I read it, much of the existing literature on imaginative resistance aims to clarify what is required for us to regard a moral claim as *radically* counterfactual in the sense required to evoke resistance.[76] I take no position on the details here, nor do I aim to provide a set of necessary or sufficient conditions for resistance to arise.[77] But I do wish to highlight the fact that the paradigmatic cases of resistance are those in which we are asked to imagine basic moral claims that we take to be radically counterfactual, and that our resistance generally disappears when we are no longer exposed to such claims; the next section will proceed under the assumption that, when such claims *are* present in a case, we have a *prima facie* reason to expect to resist it.

Recall that one outstanding question concerns whether it is possible for us to imaginatively resist thought experiments. Will we still resist a case if there is no "narrator" to whom we can attribute the moral claims we view as false? Suppose that we encounter the following variation of Infanticide in a paper on legal philosophy:

> **Jury Trial:** In killing her baby, Giselda did the right thing; after all, it was a girl. And although the jurors agreed that it was morally right for Giselda to act as she did, they also recognized that it was against the law, and they decided unanimously to convict her. Was it right for the jurors to convict Giselda, given

76 I take, for example, Walton (1994), Driver (2008), and Weatherson (2004) to be addressing this question. Walton suggests that it is conceptually impossible claims that we regard as radically counterfactual in the required sense; Driver suggests that it is psychologically impossible claims; and Weatherson suggests that it is claims which violate the dependence relationships which we believe to be actual.

77 Since my interest here is limited to our resistance to moral claims, the conditions which I propose are almost certainly not *necessary* ones for resistance – Yablo (2002) and Weatherson (2004) have argued that resistance can also emerge in response to *non-moral* claims that we regard as radically counterfactual.

that her action was illegal but morally right?

Jury Trial contains the same moral claim as Infanticide. I take it, however, that most of us will *not* experience imaginative resistance in response to Jury Trial. Why not? One possibility is that the context of a philosophical thought experiment makes resistance impossible; with no narrator to whom we can attribute the objectionable moral claim, we have a stronger impetus to "force" ourselves to imaginatively engage with it. And perhaps, when we have a reason to force ourselves to engage imaginatively with such moral claims, we are able to do so without incident.

This explanation, however, seems unlikely. What we resist, in cases like Infanticide, is not that there is a narrator who we take to be unreliable. What we resist is the *moral claim* that the lives of girls are worth less than those of boys. Our attribution of this belief *to* the narrator, in fictional vignettes, is a *manifestation* of our resistance to the moral claim rather than its cause. Because it is the objectionable claim that triggers our resistance, and because the same claim can be incorporated into thought experiments, we should expect resistance to be possible in response to thought experiments as well, even if one standard manifestation of resistance is impossible.

If resistance *is* possible in response to thought experiments, what explains why we do not experience it in response to Jury Trial? The answer is that we engage in a different sort of mental behavior when we encounter potentially problematic claims in thought experiments – a behavior which, unlike the attribution of a false belief to the narrator, can make our potential resistance disappear. When we consider a case like Jury Trial, we automatically search for and insert *qualifying contexts of our own*. Jury Trial provides no more detail than Infanticide as to why Giselda's action is right, and we are free to imagine that it is right because the lives of girls are valueless. Upon reflection, I take it, most of us will realize that we do *not* do this when considering Jury Trial. Instead, what we do is imagine that there is some other set of conditions

which makes Giselda's action right – perhaps, like in Patriarchy, there is some terrible threat faced by baby girls in the world of Jury Trial. We can succeed in imagining cases like Jury Trial without resistance, but only because our natural reaction is to insert additional conditions which make the case true *without* requiring the truth of any basic moral claims which we believe to be radically counterfactual.

The frequent availability of such qualifying contexts, as well as the ease with which professional philosophers can generally find them, explains why we rarely experience resistance in response to thought experiments and why existing work on resistance has focused primarily on fiction. But the fact that we naturally insert such qualifying contexts *rather than* engaging imaginatively with basic and radically counterfactual moral claims suggests that these claims themselves are still objectionable to us, even in the context of a thought experiment. The preceding discussion raises a question: What will happen if we encounter a problematic moral claim, but we can *neither* attribute it to the narrator of a fictional work *nor* avoid it by inserting a qualifying context? What if, for instance, we encounter a basic moral claim which we regard as radically counterfactual in the context of a thought experiment, and what if it is explicitly presented *as* a basic claim, such that we cannot insert any non-moral conditions that would make it true? I address this question in short order. For now, I turn my attention to a science-fiction tale which, we are told, bears on the debate surrounding the responsibility status of psychopaths.

III. Imaginative Resistance and Psychopathy

Levy (2007) and Shoemaker (2011) offer two versions of a thought experiment aimed at showing that psychopaths cannot be morally responsible. If their argument succeeds, then MW has a false implication. I contend, however, that their argument does *not* succeed; our intuitions in response to their thought experiment are compromised by the effects of imaginative resistance,

and therefore cannot be relied upon to support any conclusion about psychopaths. I begin by presenting Shoemaker's version of the thought experiment:

> **Grass One**: Suppose a race of alien beings comes to live among us, and while in general they share our moral sensibilities, they find additional sources of moral reasons around them. In particular, they think it immoral to walk on the grass, precisely because of what it does to the grass: it bends and breaks it. It is intrinsically bad, they claim, for this sort of organism to be bent or broken, and they purport to ground this claim on their understanding of what it is like to be a blade of broken or bent grass. When it is pointed out to them that blades of grass do not feel or have consciousness, that there is nothing it is like to be a blade of grass, they reply that understanding what it is like to be something need not have anything to do with consciousness; sometimes, it can simply consist in projectively entering into the entity's being-space. Indeed, claim the aliens, they have the special capacity for just that, and they have come to recognize the grass's moral status thereby. We, of course, simply do not get what they are talking about. Suppose, finally, that in all other physical and psychological respects, the aliens are just like us.[78]

The aliens in Grass One are supposed to stand in relation to us as we stand in relation to psychopaths. The aliens are responsive to a set of moral reasons to which we are not – the reasons pertaining to the well-being of entities like grass. And their responsiveness is due to a perceptual or imaginative capacity that we lack – the ability to "projectively enter" another entity's "being-space." Our lack of responsiveness to these particular moral reasons is intended to be analogous to a psychopath's lack of responsiveness to *any* moral reasons, and our inability to "projectively enter" the grass's "being-space" is intended to be analogous to a psychopath's lack of empathy, which is often cited as the cause of their moral defects.

Shoemaker then goes on to imagine that a human being is unmoved by the aliens' claims about the moral status of grass, and asks whether that human could appropriately be blamed for his actions:

78 From Shoemaker (2011), p.625; Shoemaker originally labels the case as "Aliens".

> What, though, of cases in which I fail to respect grass's alleged moral status? Suppose, for instance, that as I am walking through the park I see an interesting rock formation I would like to see up close but to do so involves tramping on some grass. I cannot "empathize" with the grass, and what the aliens deem immoral about grass-tramping I merely see as stupid: I am just incapable of viewing the grass's bending and breaking as giving me reasons of any kind. So as I chortle about the aliens' ridiculous moral beliefs, I tramp across the grass. I am spotted by an alien, however, who rails at me with indignation, hell-bent on publicly shaming me. Is this an appropriate reaction?[79]

Shoemaker answers this question in the negative. Because humans are unable to appreciate their moral reasons not to step on the grass, it is not appropriate for the aliens to blame them. The implication is supposed to be that psychopaths cannot be held accountable for their actions either – we cannot legitimately blame them for failing to respond to moral reasons the force of which they are unable to appreciate. The thought experiment is intended in part to illustrate certain theoretical claims about what is required to express ill-will – claims which, I argued earlier in this chapter, have already been neutralized by my arguments for MW. But as I read Shoemaker's argument, significant support is also meant to be provided by our *intuitions* about Grass One, and these are *not* neutralized by my arguments in the previous chapters. It intuitively seems that we humans would not be apt targets for blame in this case; since psychopathic bad action is relevantly analogous to our walking on the grass, Shoemaker suggests, we should be willing to accept the conclusion that psychopaths are not apt targets for blame either. My aim in the remainder of this chapter is to show that this appeal to our intuitions is unsuccessful.

I should first clarify that Shoemaker does not explicitly state that the aliens are *right* about the moral status of grass, nor does he frame his question in terms of whether humans would be blame*worthy* for stepping on it. Instead his question is whether it is appropriate for the

79 From Shoemaker (2011), p.626.

aliens to *blame* us.[80] A discussion of this question is consistent with the possibility that the aliens are mistaken about the moral status of grass, and that we are right in regarding their moral concerns as "stupid." But this would make the case disanalogous to psychopathy. We, presumably, are *right* about many of the moral claims that psychopaths view as stupid. Furthermore, our primary interest in psychopaths is in whether they are blame*worthy,* rather than in whether blam*ing* them is a practice that is rational or justified from our point of view. To make Grass One properly analogous to psychopathy, we should assume that the aliens are right, and that we really do have moral reasons not to step on the grass; from this point forward, I will understand the case in this way.

As written, Grass One is underdescribed in two critical respects. First, Shoemaker's explanation for *why* it is wrong to step on the grass is unclear. The wrongness, according to the aliens, has something to do with what it is like to be grass; and the vague notion of "being-space" is suggested as an explanation for how a non-conscious entity can nevertheless have what-it's-likeness. This vagueness is presumably intended to add to our impression that the humans in the story are unable to understand their moral reasons not to step on the grass; unfortunately, it also ensures that *we*, the readers, are unable to understand the case and thus limited in our ability to draw conclusions from it. Second, while Shoemaker indicates that the humans do not understand the aliens' moral claims, he doesn't make it clear exactly what the humans *do* believe about grass.

Talbert (2012, 2014) discusses two possible interpretations of Shoemaker's grass case, which I paraphrase here as variant cases.

> **Grass Two**: The aliens have informed humanity that stepping on grass is morally wrong; the reason for its wrongness is that grass has an inner life, and has bad experiences when stepped on. However, humans do not know about the features

80 More precisely, Shoemaker's question is whether "accountability"-type blame is appropriate.

> that *make* it wrong to step on the grass. Either the aliens have not properly explained that grass suffers when stepped on, or they have explained it but humans do not believe it. Ignorant of the features that make stepping on grass wrong, humans continue to do so.
>
> **Grass Three**: The aliens have informed humanity that stepping on grass is morally wrong; the reason for its wrongness is that grass has an inner life, and has bad experiences when stepped on. The aliens have clearly explained that grass suffers when stepped on, and the humans believe them. Nevertheless, they do not care about the bad experiences that they cause for the grass when they step on it. In full knowledge of the features that make stepping on grass wrong, humans continue to do so.

Talbert thinks that Grass Two represents what Shoemaker actually has in mind, and that Grass Three represents a distinct case. My own view is that Shoemaker's case is ambiguous, and I prefer to present Grass Two and Grass Three as disambiguations of Grass One – they are different ways of filling in the missing details of the original case. I argue shortly that the ambiguity of Grass One plays an important role in allowing the case to be processed without imaginative resistance, so the distinction between this case and the others is important. Talbert's variant cases do seem to represent the two most plausible ways of filling in the details of Shoemaker's case. If it *were* actually wrong to step on grass, this would presumably have to be because grass has an inner life and has bad experiences when stepped on; so Talbert's explanation for the wrongness of the action seems to be the only reasonable one. And, of course, humans could be either aware or unaware of grass's inner life, and the two versions of the case reflect this.

Talbert argues that the examination of these two variant cases gives us grounds to reject Shoemaker's argument. In Grass Two, Talbert notes, it seems quite plausible that humans would not be blameworthy for walking on the grass. But this is because Grass Two is a case of exculpatory non-moral ignorance – the humans are excused because they are unaware of the

features that make stepping on the grass wrong. Psychopaths typically *are* aware of the features that make their actions wrong – they know that other people have mental lives, and they know that by acting they cause pain and frustrate the preferences of others. The problem with psychopaths is that they are not *motivated* by what they know about the effects of their actions. Therefore, Talbert claims, psychopathy is more closely analogous to Grass Three, in which humans are aware of the features that make stepping on the grass wrong, but fail to be motivated by these features. The problem for Shoemaker is that it seems as though the humans in Grass Three *are* blameworthy for their actions. They know that grass has an inner life and can be harmed, and they are unmoved by this knowledge; their subsequent stepping on the grass thus seems to represent culpable *indifference* rather than potentially-excusing *ignorance*.

I agree with Talbert's attributions of blameworthiness in these cases, but I disagree with the claim that psychopaths are analogous to the humans in Grass Three. The problem with Grass Three is that the moral reason not to step on the grass is a reason that humans recognize and care about in other contexts – a morally normal human will be motivated not to cause bad experiences for some beings, such as animals and other humans. Thus the humans in this case are unlike psychopaths – they have demonstrated that they have a general ability to appreciate moral reasons of this kind, and their failure to respond to this moral reason as it relates to grass represents a local moral failing rather than a global one. I propose the following variant as an alternative:

> **Grass Four**: The aliens have informed humanity that stepping on grass is morally wrong, because it causes green things to bend, and bending green things is intrinsically and irreducibly morally bad. Bending green things, as it turns out, is simply one of the items on the list of features that ultimately make actions wrong, alongside such others as violating a duty or causing suffering. The aliens' moral claims are correct, and it really is morally bad to bend green things. Humans recognize that grass is green and that it bends when stepped on; they remain

unmotivated by this knowledge, however, and continue to step on the grass.

Grass Four avoids the problem facing Grass Three. There is *no* context in which normal humans recognize a moral reason not to bend objects with a particular color; thus, the humans in this case may reasonably be understood to be *incapable* of appreciating this kind of moral reason. If we want to use a thought experiment as an analogy for psychopathy, then it is a case like Grass Four that we should use. So we might pose Shoemaker's question again, this time with respect to Grass Four – assuming that it really is intrinsically and irreducibly morally bad to bend green things, and humans know that they bend green things when they step on the grass, does it seem as though the humans are blameworthy?

My own sense is that the humans in this case would not be blameworthy. Others may share this intuition, or have intuitions that differ. My central contention in this section, however, is that *our intuitions about Grass Four do not matter.* That is, we have good reason to think that our intuitions are *not* a reliable guide to whether the humans in this case would *actually* be blameworthy. Grass Four, I contend, is a case that we cannot or will not properly imagine due to imaginative resistance. At best, we are willing and able to imagine a modified case, lacking the central moral claim – that it is intrinsically and irreducibly morally wrong to bend green things. Since the modified case is the one that we imagine, it is the one that we form intuitions about; since the modified case will not be analogous to psychopathy, our intuitions about it are useless as evidence for the blameworthiness, or lack thereof, of psychopaths.

Such, at least, is the outline of my argument. In the remainder of this chapter, I will work through the argument more slowly. I begin with the first claim – why should we think that we experience imaginative resistance in response to Grass Four? We are unlikely to have any strong sense that the author lacks the authority to make his moral claims true, as we do in response to

literary cases that we resist. But this is most likely due to the conventions governing philosophical thought experiments, according to which the author is free to stipulate any claim as a hypothetical. I suspect that most of us *will* experience the other main phenomenological indicator of resistance when we consider this case – if we try to imagine a world in which bending green things is intrinsically and irreducibly wrong, we are likely to have the sense that we do not really succeed in doing so. Grass Four also seems to satisfy the conditions for resistance described in the previous section. It presents a moral claim which is *basic* – the assertion that bending green things is intrinsically morally bad – and which most of us will regard as radically counterfactual. These were admittedly not presented as necessary or sufficient conditions for resistance to emerge, but their satisfaction does seem to provide us with a *prima facie* reason to think that we will resist Grass Four.[81]

The history of the case itself provides additional evidence, as it demonstrates the avoidance behaviors that we should expect in response to a claim that we resist. Grass Four, I argued, is the only version of the case which is properly analogous to psychopathy. Yet neither of the previous discussions employed Grass Four; instead, each earlier iteration of the case incorporated a qualifying context that rendered the case disanalogous to psychopathy but also prevented us from having to imagine any basic, radically counterfactual moral claims. We do not resist Grass One, but this is because the case offers an "explanation" of sorts for the wrongness of stepping on the grass – it has something to do, the aliens tell us, with "being-space." I regard the appeal to being-space in Grass One as a kind of qualifying context; although not illuminating,

81 By way of objection, one might point out that many of the prototypical cases of imaginative resistance (e.g. Infanticide, First Contact) incorporate moral claims that we find emotionally *repugnant*, and that we are likely to find the moral claim in Grass Four baffling rather than repugnant. It seems, however, that resistance does not require a strong emotional response to be triggered – it can occur in response to emotionally neutral and even non-moral claims (Yablo 2002, Weatherson 2004).

it draws our attention away from any basic moral claims and thus prevents us from tying to imagine them.[82] Grass Two and Three incorporate a different qualifying context – they stipulate that grass can suffer, which allows us to imagine that stepping on it is wrong without changing the basic moral facts that are true in the actual world. Although Grass Four is the case that we should be considering, the earlier iterations seem to reflect automatic efforts to *avoid it*; this is strongly suggestive of a case that triggers resistance.

So I think that we do experience imaginative resistance in response to Grass Four. But why should this resistance interfere with our ability to form reliable intuitions about the case? Here is what I propose happens. Although we attempt to form intuitions about Grass Four, imaginative resistance prevents us from successfully doing so. Because we either cannot or will not imagine that bending green things is intrinsically morally wrong, we non-consciously *omit* this claim from the case that we consider. The case that we actually imagine is therefore a "censored" version, and is not really Grass Four at all – call the censored version Grass Four*. It is Grass Four*, rather than the original Grass Four, that is fed into whatever cognitive mechanisms process cases in order to produce moral intuitions. And because Grass Four* lacks the moral claim that is necessary to make Grass Four analogous to psychopathy, the intuitions we form about it are of no use for Levy and Shoemaker's purposes – whatever intuitions we form in response to Grass Four* have no bearing on the blameworthiness of psychopaths.

For the sake of clarity, it would be helpful to provide a more detailed description of the content of Grass Four*, perhaps offering a precise characterization as a variant case. Unfortunately, this is not possible. Crucial to my account here is the fact that the modification

82 See Gendler (2000), p.75, for a discussion of qualifying contexts which prevent resistance in a similar way, drawing our attention away from potentially objectionable moral claims rather than explaining them.

does not occur consciously. We still *believe* ourselves to be considering Grass Four, even though we are not; thus, we lack introspective access to the content of Grass Four*, and cannot precisely describe it. To an approximation, Grass Four* may resemble any one of the earlier cases, Grass One through Three; its content may even vary across individuals. But the essential feature of Grass Four* is that it lacks the problematic moral claim – so it is *not,* in the world of Grass Four*, true that bending green things is intrinsically and irreducibly morally wrong. The lack of this moral claim is sufficient to make the case disanalogous to psychopathy, so my argument does not depend on our knowing exactly what this claim is replaced with.

We might first ask whether the story I have proposed is even possible – can an agent really be *mistaken* as to which case she is forming intuitions about? In order for this to occur, it seems, our moral intuitions must be generated at least partly by a process or faculty to which we lack direct conscious access. We need not endorse any particular account of moral cognition, however, to conclude that the general view of intuition formation required for my hypothesis is a very plausible one. It is supported by the emerging consensus that the mind is modular to at least a significant degree[83], as well as our developing understanding of the importance of non-conscious processes in a variety of cognitive roles.[84] More specifically, the phenomenon of moral dumbfounding seems to demonstrate fairly decisively that intuition-generation is handled at least in part by non-conscious processes[85]: Agents often report having moral intuitions about cases without being able to explain why, which strongly suggests that these intuitions are produced by a process that is not generally available for conscious access. Because this process is not consciously accessible, it is possible for us to be mistaken about which case we are forming

83 See, e.g. Robbins (2009).
84 See, e.g. Libet (1985), Wegner (2002), Carruthers (2011), Levy (2015).
85 See, e.g. Haidt (2001).

intuitions about.

So much for this worry – the story I propose here should at least be *possible*. But why think that it actually occurs? That is, why think that Grass Four actually *is* censored into Grass Four*? One reason is that this would seem to be a likely consequence of the effects of imaginative resistance discussed in the previous section. When we resist imagining a claim, we are often left with the sense that we fail to imagine it; as claimed earlier, this would seem to provide at least *prima facie* evidence that we do *not* imagine the claims that we resist, and that, in the world we are imagining, the resisted claim really *is* false.[86] And this in turn would seem to provide evidence that the scenario sent to our moral faculties for processing is one which lacks the problematic moral claim.

The non-conscious censorship of problematic cases also seems to be a likely consequence of our automatic efforts to avoid imaginatively engaging with those claims that we resist. When we resist a fictional case, we are free to do so by attributing the objectionable moral claims to the narrator – whoever is describing the fictional world, we think, must be wrong. This reaction is not available in response to thought experiments, so we instead search for qualifying contexts that could make the problematic moral claim true. But what happens when there is no *room* for such qualifying contexts? What if the basic moral claim is explicitly described as such, and we cannot imagine its truth without imagining a radical change to the moral facts which we believe to be actual? One possibility is that our mental efforts to avoid imagining the claim will cease. But this seems unlikely, since, as argued in the previous section, the claim itself will not have changed and will remain as objectionable as ever. It seems much more likely that we will

86 To be clear: I do not mean to make any controversial assumptions about the way truth works in fictional worlds, or even to assume that there *is* such a thing as truth in fictional worlds. When I say that P *is* false *in* the world we are imagining, what I mean is that we are, in trying to imagine that world, representing to ourselves a world in which P is false.

continue to attempt to avoid engaging with the problematic moral claim. And in such a case, it seems that we have only one option for avoidance – we must drop the problematic claim entirely, failing or refusing to actually imagine it. This is exactly what I propose happens. Because we cannot or will not accept the claim that bending green things is intrinsically and irreducibly morally wrong, we non-consciously omit it from the case. And the new, revised case, lacking the moral claim of interest, is Grass Four*.

To recap: What I propose is that we do not really imagine Grass Four, even when we believe ourselves to be doing so. Instead, we imagine a superficially related case, Grass Four*, which lacks the problematic moral claim. But this missing moral claim would be necessary to make the case properly analogous to psychopathy. So, even if our intuitions indicate that the humans in this case would not be blameworthy, they give us no reason to think that psychopaths are not blameworthy. The case that we are actually imagining is not like psychopathy, and we cannot assume that our intuitions – even if a reliable guide to truth in Grass Four* – provide us with any insight into the blameworthiness of psychopaths.

The above does not, of course, establish that psychopaths *are* blameworthy – that was one major goal of the preceding chapters. What I have done in this chapter, however, is to show that a particular argument against the blameworthiness of psychopaths – in my opinion, the strongest argument – fails. Thus, I conclude that we should accept MW's implication that psychopaths are blameworthy, rather than regarding it as a *reductio*.

Chapter Seven

Positive and Negative Moral Incompetence

I. Introduction

On AG accounts, moral virtue consists in having the appropriate attitudes towards the features that actually make actions right or wrong, rather than towards the the good or the bad *de dicto*. Thus these accounts imply that agents who fail to display the right attitudes towards the right- and wrong-making features also fail to display virtue, irrespective of whether these agents believe themselves to be acting rightly. MW, the account defended in previous chapters, refines this picture, as it allows us to determine precisely *which* features of actions are right- or wrong-making in the relevant sense. Specifically, according to MW, the features of interest are those that make actions right or wrong by performing normative explanatory work at the lowest level. I argued in Chapter Five that, in addition to resolving the problem cases, MW allows us to evaluate many agents who are mistaken about the permissibility of controversial practices such as abortion. In many such cases, it turns out that moral disagreements are due to differing attitudes towards an agreed-upon set of low-level features, rather than to disagreements as to which low-level features are present. One implication is that many agents who turn out to be wrong about such practices fail to display the correct attitudes towards the low-level features, and thus do not display virtue – despite the fact that they may appear "well-meaning."

In the case of practices that turn out to be impermissible, those agents who fail to possess the appropriate anti-attitudes towards the low-level wrong-making features are vicious, and are blameworthy when their lack of appropriate attitudes is reflected in their actions. In the preceding discussion, however, I left open the question of how we should evaluate those cases in

which the disputed practice turns out to be *permissible*. What should we say about agents who form anti-attitudes towards features of actions that are *neither* right- nor wrong-making, but morally irrelevant? These agents are *not* virtuous, I argued, since their attitudes are not good. But it may seem that they cannot be *vicious*, either. After all, the features towards which they display anti-attitudes are morally neutral, and it may seem that the attitudes themselves must be morally neutral as well.

I revisit this question in the current chapter, and argue that agents of this kind *are* morally vicious. This result cannot be produced either by MW in its current form or by other existing AG accounts, and will require them to be augmented. The required augmentation is, however, a theoretically plausible one. On AG accounts, moral virtue is understood to be a kind of competence at responding to moral reasons, or, alternatively, at responding to the considerations or features of actions that provide moral reasons. The conception of moral competence at work in existing AG accounts seems to be one on which perfect competence consists simply in responsiveness to *all* moral reasons, or to all of the considerations or features of actions which provide them. I argue that this conception should be expanded: We should understand perfect moral competence to consist in responsiveness to all *and only* the actual moral reasons, or to all *and only* the features or considerations that actually provide such reasons. This more expansive conception of moral competence includes the previous, narrower one as a special case. But it additionally implies that agents are morally worse in virtue of their responsiveness to morally irrelevant considerations as though they provide moral reasons. This implication not only allows us to confidently evaluate agents in cases of moral disagreement, but also allows us to explain the blameworthiness of a certain class of wrongdoers who are otherwise difficult to make sense of.

In the next section, I motivate the discussion in this chapter by describing a set of cases in which agents intuitively seem to be blameworthy, but which neither MW nor existing AG accounts are able to explain. In the third section, I describe my proposed augmentation, which requires us to adopt a more expansive account of moral competence. In the fourth and final section, I discuss some implications and offer some remarks intended to conclude both this chapter and this dissertation.

II. Wrongdoers Who Care About the Actual Good

Elizabeth Harman (2011) notes a puzzle posed by agents who satisfy the following four criteria:

> (a) [The agent] acts wrongly while believing a false claim, p,
> (b) if p is true then the action is permissible,
> (c) the false belief did not result from mismanagement of belief, and
> (d) the false belief is not a case of motivated ignorance[.][87]

Not only do agents who satisfy these criteria often seem to be blameworthy, Harman notes, but they may account for many of the most interesting and morally important cases of blameworthiness; many war criminals and ideologically-motivated terrorists, for instance, may satisfy these criteria. Yet it is difficult to explain *how* such agents could be blameworthy for their actions. By stipulation, they believe themselves to be acting rightly. And since their false beliefs are sincere and formed without either epistemic mismanagement or self-deception, it would seem that we cannot trace their blameworthiness back to an instance of negligence or other epistemic bad action. In virtue of *what* could these agents be blameworthy?

Harman suggests that we can answer this question by appealing to an AG account like Arpaly's (2003). She points out that agents who satisfy these criteria may nevertheless

87 Criteria reproduced from Harman (2011), pp.455-6.

demonstrate inappropriate attitudes towards that which is morally important when they act. Specifically, she proposes, these attitudes may be reflected in the false moral *beliefs* that produce the bad action. A racist belief, even though acquired without motivated ignorance or epistemic mismanagement, is nevertheless likely to reflect some kind of inappropriate attitude towards that which actually good and bad – a lack of concern for the rights-conferring properties of people of other races, for instance. Because we can appeal to these faulty attitudes, we *can* explain why these agents are blameworthy.

Harman's proposed solution prefigures the one defended earlier in this dissertation. Her proposal is offered as a defense against Rosen's (2004) skeptical challenge to moral responsibility; in Chapter One, I argued that the ability of AG accounts to defeat this skeptical challenge is a major consideration in their favor. Thus, I think that Harman's solution is essentially correct, and that it works successfully for many agents.[88] The problem is that it will *not* work for *all* of the agents who strike us as obviously blameworthy.

Consider an agent who satisfies all of criteria (a) through (d) above, but additionally satisfies the following criterion:

> (e) neither the wrong action nor the false belief p reflects a lack of concern for anything morally important.

Can we imagine agents who satisfy all five criteria, including (e)? We can. Consider the following variation on Clinic Bomber:

> **Clinic Bomber Plus:** George is deeply concerned with the well-being of persons, and strongly desires to prevent persons from being killed. Because he believes that fetuses are persons, he believes that he can save persons by preventing abortions. Accordingly, he places a small bomb in an abortion clinic and detonates it at a time when he knows the clinic will be unoccupied. The resulting damage to the facility,

88 Unlike Harman, I see no need to appeal to an agent's *beliefs* in most of these cases, and am happy simply to describe the bad attitudes as being reflected in the action directly, without requiring that they be reflected indirectly by way of the beliefs.

> which is located in an area with limited access to abortion, forces it to close for several weeks and prevents a number of abortions which would otherwise have taken place.
>
> George originally became convinced that fetuses were persons when he read a description of their biology. A fetus has a complete and unique human genome, and this, George thinks, endows it with personhood. This belief is false – a being actually requires certain psychological properties in order to be a person – but George came to acquire it by reasoning responsibly and without self-deception. Fetuses do not in fact have the required psychological properties, so George's action does not actually save any persons. Even so, saving persons *is* morally good, and, if fetuses *were* persons, George's action would have been morally right.
>
> George knows that his action will harm women and violate the property rights of others. Even so, George is not *lacking* in concern for property rights or for women's well-being. He cares exactly as much about these things as he should, and normally acts so as to promote them – he would be willing to make a significant personal sacrifice, for example, to protect the well-being of women. His concern for the well-being of fetuses is simply so great that it overpowers his concern for these other considerations.

Clinic Bomber Plus is identical to the previous presentation of this case, except for the final paragraph, which has been added to make it clear that George satisfies condition (e). George cannot display any inappropriate attitudes towards that which is actually morally important, because he does not *have* any such attitudes – he is stipulated to care exactly as much as he should about everything that matters morally. It therefore follows that neither Harman's solution nor existing AG accounts more generally can explain why George is blameworthy. Blameworthiness, on these accounts, requires the expression of attitudes which George is stipulated to lack.

My main task in this chapter is to describe how AG accounts can be augmented to produce the result that agents like George *are* blameworthy, but it will first be important to establish that this result is a desirable one. I originally presented Clinic Bomber alongside other "problem cases", which were all offered as cases in which it was unclear how we should evaluate

an agent. I am now presenting Clinic Bomber Plus as an example of a case in which it seems clear that an agent *is* blameworthy, in order to motivate the theoretical changes needed to accommodate this judgment of blameworthiness. Is this rhetorical shift a legitimate one? I believe that it is. The correct reaction to the original Clinic Bomber was unclear, I argued, because there was some reason to believe that George is actually praiseworthy. After all, his action reflects a desire to save persons, which was stipulated to be morally good. I argued in previous chapters, however, that the desire to save persons does *not* count as a good desire for the purposes of assessing moral worth. In light of these arguments, any reason we might once have had to regard George as praiseworthy has been eliminated. And in the absence of such a reason, I contend, it should strike us as fairly obvious that George is blameworthy.

I do not mean to claim that this view of George is *inevitable*. I have previously discussed Rosen's (2004) view, on which "clear-eyed akrasia" is required for blameworthiness; this view will have the consequence of excusing George as well as a wide range of other agents whom we are intuitively inclined to blame. But while there are various accounts that would compel us to *deny* that agents like George are blameworthy, I think this denial would be significantly *revisionary*. As noted in Chapter One, I consider certain judgments about blameworthiness to lie close to the foundations of our moral thinking. Many agents like George are likely to be among those whom we have a strong intuitive tendency to judge blameworthy; if an account of moral worth can accommodate this judgment, then this should be a consideration in its favor.[89]

I argue in the next section that AG accounts *can* be modified, in a theoretically plausible way, to accommodate this judgment. One might worry, however, there there are *other* ways of

89 If needed, we can elicit stronger intuitions by modifying the case to describe a more dramatic instance of wrongdoing – perhaps George is a war criminal who cares deeply about the people he kills, but who cares even *more* about ethnic homogeneity.

accommodating this judgment that would not require us to modify our account of moral worth. I discuss two alternatives here, both of which are ways of maintaining that George really *does* express inappropriate attitudes towards that which is morally important. The first locates these in his implicit attitudes towards considerations unrelated to personhood – such as property rights or the autonomy of women – while the second locates these in his implicit attitudes towards the actual personhood-conferring psychological properties.

I stipulated in Clinic Bomber Plus that George cares exactly as much as he should about everything that is actually morally important. One might first object that this is impossible given the other stipulations made in the case. George clearly cares more about fetuses than he does either about women's autonomy or about property rights. Why not think that this is *constitutive* of insufficient concern for these other considerations? That is, why not think that to care less about the autonomy of women than about fetuses simply *is* to care insufficiently about the autonomy of women?[90] This alternative is most plausible if we understand the strengths of an agent's attitudes to be relative rather than absolute. Were the strengths absolute, we could simply stipulate that an agent ought to have, say, one hundred units of concern for women's autonomy, and further stipulate that George has precisely this amount. If, in contrast, the strengths are relative, we can only say that an agent ought to care *more* about women's autonomy than about certain other things. By caring *less* about women's autonomy than the genetic properties of fetuses, one might think, George cares less about it than he should.

But even if the strengths of an agent's attitudes are relative rather than absolute, this alternative seems unlikely to succeed. The reason is that a *single* comparison cannot be sufficient to establish the strength of an attitude. George cares less about women's autonomy than he does

90 Many thanks to Yishai Cohen and Julia Markovits for independently raising this question.

about fetuses. But it is important to remember that George cares *more* about women's autonomy than he does about many other considerations. He cares more about it than his own well-being, for example – I stipulated that George would make significant personal sacrifices in order to protect women's autonomy. We can imagine that if these pairwise comparisons were repeated, George would care more about women's autonomy than *any* other consideration, *except* for the well-being of fetuses. It seems implausible to describe an agent with these attitudes as lacking in concern for women's autonomy; the same goes, *mutatis mutandis*, for property rights, or for any other consideration that we think might be relevant.[91]

The second alternative is to maintain that George *does* display inappropriate attitudes towards the actual personhood-conferring properties. George thinks that to be a person is to have certain genetic properties; we might claim that, in so thinking, George must also implicitly think that psychological properties are irrelevant to personhood and therefore morally unimportant. So perhaps his concern for the genetic properties is part-and-parcel with a lack of concern for the psychological ones, and perhaps his display of the former also constitutes a display of the latter.

But this alternative also seems unlikely to succeed. We could easily modify the case so that George cares about the genetic properties *as well as* the the psychological properties which actually confer personhood. Perhaps he thinks that the conditions for personhood are disjunctive, and that a being with either the genetic or the psychological properties is a person. We can

91 If the preceding discussion seems insufficient to show that this alternative does not work, we can imagine that George starts out with a *perfect* set of attitudes – he cares exactly as much as he should about everything, which includes his not caring at all about fetuses. He then *comes* to care a great deal about fetuses after he is exposed to certain arguments for their moral status, without experiencing any other psychological changes. It does not seem as though we should describe the change that George undergoes as his coming to care *less* about everything *but* fetuses than he did previously; rather, the change is simply that he has come to care *more* about fetuses. Alternatively, imagine another morally perfect agent who suddenly stops caring about whether he tells the truth, without undergoing any other psychological changes. For every consideration that this agent actually cares about, it will now be the case that he cares about it more than he cares about telling the truth. But it does not seem apt to describe this agent as caring *too much* about everything *but* telling the truth; rather, his problem is simply that he cares *too little* about telling the truth.

further imagine that he is just as strongly motivated to protect beings with the psychological properties as he is to protect beings with the genetic ones. Presumably we would still judge George blameworthy in this modified case, and yet his blameworthiness cannot be traced to a lack of concern for the actual personhood-conferring features. One might maintain that caring *as much* about genetic properties as about psychological ones somehow *constitutes* caring too little about the psychological properties. But this would be analogous to the approach described above – on which we understand caring about fetuses to be constitutive of caring too little about women – and seems implausible for the same reasons. George could care much *more* strongly about the psychological properties than about many other considerations, and might be willing to make personal sacrifices in order to protect those beings that possess them; the fact that he *also* cares about an additional set of properties does not seem to diminish his concern for the psychological ones.

III. Positive and Negative Moral Incompetence

Having rejected these alternatives, it seems that the only way of accommodating our intuition that agents like George are blameworthy is to expand our conception of moral competence. This expansion allows for the possibility of two kinds of moral *in*competence, in two "directions." One, negative moral incompetence, is familiar from previous examples – this is the kind of incompetence displayed by agents who fail to have the correct attitudes towards the actual good and bad, and it is already accommodated by existing AG accounts. The other, positive moral incompetence, is new – it is the kind of moral incompetence displayed by agents like George, who respond to irrelevant features *as though* they provide moral reasons. Both kinds of incompetence, I argue, represent deviations from a moral ideal – the ideal of responsiveness to all *and only* those considerations that are actually morally important – and the more expansive

conception of moral competence that I propose here is intended to capture this ideal. Because both kinds of incompetence represent deviations from virtue, both are vicious, and we can appeal to the vice represented by positive moral incompetence to explain why agents like George are blameworthy.

Formally, the relatively narrow account of moral competence which I believe to be implicitly at work in existing AG accounts is as follows:

> Narrow Moral Competence: An agent's moral competence is a measure of whether he responds in a morally appropriate way to the features of actions that actually provide moral reasons. Perfect moral competence consists in responsiveness to all features that actually make actions right or wrong.

I propose that we adopt the following, more expansive account of moral competence:

> Expansive Moral Competence: An agent's moral competence is a measure of whether he responds in a morally appropriate way to the features of actions. Perfect moral competence consists in responsiveness to all *and only* those features that make actions right or wrong.

It will first be helpful to say a bit more about why the move from the narrow conception of moral competence to the expansive one should seem theoretically plausible. I begin with an appeal to Aristotle. While AG accounts are not Aristotelian, the Aristotelian view is nevertheless useful in that it illustrates some apparent truisms about virtue and vice. Suppose that *charity* is a stereotypically Aristotelian virtue and that it is a measure of whether an agent is appropriately motivated to give his money to others who are less fortunate.[92] One way of failing to be charitable is to *fail* to give money to those to whom one ought to give. Naïvely, we might suppose that perfect charity consists in *always* giving money when the opportunity arises; alternatively, we might suppose that there is some subset of occasions, such that a perfectly

92 My description of "charity" is meant to be similar to what Aristotle refers to as "liberality" in "the giving and taking of money"; see e.g. *Nicomachean Ethics* Bk.II Pt.7. Whether "charity" corresponds exactly to a virtue described by Aristotle is irrelevant, so long as it is understood to be Aristotelian in form.

charitable agent would give on *at least* those occasions. But neither of these, of course, is the Aristotelian view. Virtues like charity admit of defects in two directions, and it is possible to be vicious both by giving on *too few* occasions and by giving on *too many.* An agent cannot simply give on every possible occasion and be virtuous, for an agent who gives too often is prodigal. The maximally charitable agent is one who gives on all *and only* those occasions on which it is appropriate to do so.

To reiterate, AG accounts are not Aristotelian, and these remarks are less about Aristotle than they are about our intuitive commitments concerning the nature of virtue. One such commitment seems to be the following: The virtuous person is not someone who is motivated in a certain way to the maximum degree, or who acts in a certain way on every occasion. Instead, the virtuous person is one who is motivated in a certain way to a certain degree and *to that degree only,* or a person who acts in a certain way when *and only when* it is appropriate. One might nevertheless question whether these observations are of any relevance to the current discussion. After all, Aristotle's conception of virtue is much broader than the one that interests us here; his use of "virtue" encompasses many forms of human excellence, whereas the sole concern of AG accounts is *moral* virtue.

But it is important to remember that AG accounts commit us to a model on which moral virtue is analogous in important ways to competences in other domains. Insofar as these domain-specific competences resemble Aristotle's virtues, *moral* competence – which for our purposes is equivalent to moral virtue – will also resemble them. Significantly, domain-specific competences *do* seem to resemble Aristotle's virtues in a very general way, in that they often admit of deficiencies in two directions. Let competence with respect to domain X consist in X-appropriate responsiveness to the considerations that an agent encounters while acting in his capacity as an

X-agent. An agent can fail to be competent with respect to X by ignoring or failing to respond to some of the considerations that provide X-related reasons. But she can also fail to be competent with respect to X by responding to irrelevant features *as though* they provide X-related reasons.

By way of illustration, suppose again that the *medical competence* of a doctor consists in medically-appropriate responsiveness to the features of her patients. If her patient has a condition that contraindicates a certain kind of medication, a competent doctor will respond by not prescribing that medication. If her patient has a symptom that requires a certain kind of test, she will respond by ordering that test. The failure to respond to either of these kinds of features, both of which actually provide medical reasons, would represent a defect in the doctor's quality *qua* doctor. Call defects of this kind instances of *negative medical incompetence*.

Significantly, however, there is another way for a doctor to fail to respond appropriately to the features of her patients: She could respond to features that are medically irrelevant as *though they do* provide medical reasons. Suppose that a doctor systematically responds to a symptom like abdominal pain by administering an irrelevant procedure, such as a knee-reflex test, which the symptom gives her no medical reason to perform.[93] This doctor surely also demonstrates a defect in the quality of her doctoring. The defect she displays does not consist in a failure to respond to any feature that actually provides medical reasons, but rather in *responsiveness* to a feature that does *not* provide medical reasons. Call this defect an instance of *positive medical incompetence*.

If my analysis of these medical examples is correct, then we must understand perfect medical competence to consist in responsiveness to all *and only* those features of patients that

93 We can assume that in addition to the knee-reflex test, the doctor also performs any procedures that actually *are* warranted by the patient's symptoms; this way, the case is not *also* an example of negative medical incompetence.

actually provide medical reasons. My contention is that we should understand *moral* competence similarly. Moral competence is a measure of whether an agent responds in a morally appropriate way to the features of actions that he encounters when acting in his capacity as a moral agent.[94] Perfect moral competence consists in responsiveness to all *and only* those features which provide moral reasons – as opposed to responsiveness merely to *all* the features which provide moral reasons.

Just as the more expansive account of medical competence allows for two varieties of medical *in*competence, the more expansive account of moral competence will allow for deviations in two "directions." An agent can fail to respond to some features which actually provide moral reasons; call this *negative moral incompetence*. An agent can also *respond* to some features which do *not* provide moral reasons *as though* they do; call this *positive moral incompetence*. Since both kinds of incompetence are deviations from the virtuous ideal of moral competence, both are vicious. And by appealing to positive moral incompetence, we can explain why George is vicious – he responds to the genetic properties of fetuses as though they provide moral reasons, when in fact they do not. Since this vice is reflected in George's action, he is blameworthy. Thus, the move to the more expansive account of moral competence allows us to accommodate the judgment that agents like George are blameworthy. The same analysis can be repeated for other agents of this kind – war criminals, ideologically-motivated terrorists, and so on – and thus the puzzle described in the previous section is resolved.

So the move to this more expansive conception of moral competence seems theoretically plausible, and also allows us to resolve the puzzle cases concerning agents like George; I

94 Except in special circumstances such as incapacitation, I assume that moral agents are *always* acting in their capacity as moral agents.

therefore think we have good reason to make it. The move to the expansive conception does, however, present a problem of its own: On this conception, agents are vicious when they respond to irrelevant features *as though* they provide moral reasons, and we will need an account of what it means to respond to a feature in this distinctively moral way. We cannot simply include *any* kind of reaction to a morally irrelevant feature as an instance of positive incompetence, for reactions to such features are extremely common and generally do not reflect negatively on our moral character. I may form a pro-attitude towards heading to the break room, in response to the fact that this action is a way of obtaining coffee. My reaction to the prospect of coffee is clearly not a *moral* one, and it does not demonstrate any defect in my *moral* responsiveness to the world around me. Positive moral incompetence is only intended to encompass responsiveness of a distinctively moral kind; this is what makes it plausible as a kind of defect in an agent's *moral* character.[95]

The question of precisely what should count as moral responsiveness may be a difficult one. But while a complete answer to this question would be necessary to determine the precise extension of positive moral incompetence, it is not required for the relatively narrow goal of this chapter, which is to propose a modification to AG accounts that will adequately accommodate agents like George. A full account of the necessary and sufficient conditions for moral responsiveness is not necessary here because there seem to be conditions that are clearly *sufficient* for moral responsiveness, and it seems that the agents of interest to us satisfy them. If

95 To return to the medical analogy, a doctor can respond to all sorts of medically irrelevant considerations without displaying medical incompetence, so long as she does not respond to them *as though* they are medically important. Pediatricians sometimes give lollipops to children who were well-behaved during their medical exams. If a doctor does this for business reasons or for humanitarian reasons, then her pro-attitudes towards giving lollipops would not be an instance of medical incompetence. If she for some reason thought that it was medically important for children to have lollipops after being examined, these attitudes *would* be an instance of medical incompetence.

an agent explicitly believes that a particular feature makes actions morally good, and forms a pro-attitude towards actions with that feature as the non-deviant result of that belief, then his responsiveness seems to be an obvious case of *moral* responsiveness – it clearly reflects something about the agent's competence *qua* moral agent, which is what we are after. Agents like George satisfy this sufficient condition. We can propose a similar sufficient condition that will be satisfied by agents who falsely believe certain actions to be *impermissible* – an agent who explicitly believes that a certain feature makes actions morally bad, and whose belief non-deviantly causes him to form an anti-attitude towards actions with that feature, also displays *moral* responsiveness.

So, in the context of this project, we can restrict positive moral incompetence to those agents whose attitudes are due to their false moral beliefs – the agents in the cases of interest satisfy this requirement. In imposing this restriction provisionally, I mean to leave open the possibility that the true extension of positive moral incompetence may be greater than this. Perhaps there is room for a broader notion of distinctively moral responsiveness, such that agents can display moral responsiveness to irrelevant features even without explicit moral beliefs; if so, these agents will also be vicious. As I will discuss in the next section, the main consequence of acknowledging positive moral incompetence as a vice is that we will need to acknowledge more agents as vicious and blameworthy than would otherwise be the case. If it turns out that positive moral incompetence is even more widespread than I have supposed here, this will have the effect of making the project's implications more rather than less dramatic.

Thus far, I have argued that AG theorists should adopt a more expansive account of moral competence. I have not yet commented on how this change would affect the specific criteria for moral worth offered by any AG account. The details, of course, will depend on the account.

While I think the change can be accommodated by AG accounts more generally, I focus here on MW, the account which I defended in previous chapters. Here is how I propose to amend MW:

MW (Expanded): Agents are morally *praiseworthy* for free actions that reflect one or more of the following:

a.) concern for, a motivation to promote, or an otherwise appropriate pro-attitude towards the features of actions that make them right or good and which perform the lowest-level of normative explanatory work;

b.) abhorrence for, a motivation to discourage, or an otherwise appropriate anti-attitude towards the features of actions that make them wrong or bad and which perform the lowest level of normative explanatory work; or

c.) a lack of concern for, a lack of motivation to promote, or a lack of other inappropriate pro-attitudes towards the features of actions that make them wrong or bad and which perform the lowest level of normative explanatory work.

Agents are morally *blameworthy* for free actions that reflect one or more of the following:

a.) a lack of concern for, a lack of motivation to promote, or a lack of other appropriate pro-attitudes towards the features of actions that make them right or good and which perform the lowest level of normative explanatory work;

b.) contempt for, a motivation to discourage, or another inappropriate anti-attitude towards the features of actions that make them right or good and which perform the lowest level of normative explanatory work;

c.) concern for, a motivation to promote, or an otherwise inappropriate pro-attitude towards the features of actions that make them wrong or bad and which perform the lowest level of normative explanatory work;

d.) concern for, a motivation to pursue, or another pro-attitude towards features of actions that are morally irrelevant, but to which the agent responds *as though* they make actions right or good; or

e.) contempt for, a motivation to discourage, or another anti-attitude towards features of actions that are morally irrelevant, but to which the agent responds *as though* they make actions wrong or bad.

Conditions (d) and (e) for blameworthiness have been added; the account is otherwise identical to my previous description of MW. This revision to MW accounts for positive moral

incompetence by attributing blameworthiness to those agents who morally respond to irrelevant features by forming either pro- or anti-attitudes. The conditions for moral responsiveness are not specified here; provisionally, moral responsiveness can be understood as forming the pro-or anti-attitude as the non-deviant result of an explicit moral belief. As noted, however, I leave open the possibility of a more expansive conception of moral responsiveness, and one can be substituted into MW if desired.

I conclude this section by addressing one objection: We might worry that this expanded version of MW would attribute blameworthiness to too many agents. Suppose we agree that agents like George are blameworthy, and that the appeal to positive moral incompetence is useful in its ability to explain their blameworthiness. Even so, there might be other positively incompetent agents who do *not* seem blameworthy to us. For instance:

> **Would-Be Farmer:** Susan believes that it is morally bad to kill any animal, including small invertebrates such as worms and insects. Her belief is not due to any mistake about the non-moral features of small invertebrates; she does not, for instance, mistakenly believe that worms have complex inner lives. She simply thinks that the features that small invertebrates *do* have – such as their capacity for independent movement and reproduction – make it morally bad to kill them. Because the cultivation of fields causes the deaths of many small invertebrates, Susan believes that she has a moral reason not to become a farmer. As a result, she ultimately decides not to take up a career in agriculture, something which she would otherwise have pursued. Suppose that, given the actual moral facts, Susan has no reason to refrain from killing small invertebrates; in actuality, she has no moral reasons either to become a farmer or to refrain from becoming a farmer.

Would-Be-Farmer seems to offer a fairly clear case of positive moral incompetence. The features which Susan attributes to insects are wholly morally irrelevant, yet Susan responds to them *as though* they provide strong reasons not to become a farmer. When Susan decides not to be a farmer, in light of her supposed moral reasons not to do so, her positive moral incompetence is reflected in her action. It seems that the expanded version of MW defended in this chapter

should imply that Susan is blameworthy. But this may be a surprising implication to many of us, given that Susan's action is entirely harmless.

Even more surprisingly, my claims in this chapter may seem to imply that some agents who act *well* are at least partially blameworthy. Recall Elaine, from Torture. Elaine wants to avoid causing pain; in the world of Torture, this is a low-level wrong-making feature of actions, so Elaine's aversion to causing pain is virtuous and she is praiseworthy when it is reflected in her action. But Elaine *also* wants to maximize utility. The correct normative theory, in Torture, is stipulated to be a Kantian one on which utility is irrelevant. So it seems that Elaine is *also* responsive to a consideration that is morally irrelevant, and therefore that she is at least partially vicious. Since Elaine's concern for utility is reflected in her action, it seems that she is also partially *blameworthy.* Are these implications that we should accept?

It is important, first of all, to distinguish the implication that these agents are at least somewhat *vicious* from the implication that they are *blameworthy.* I think that we certainly *should* accept the former implication. Both Susan and Elaine are morally responsive to something that is morally unimportant; this responsiveness represents a defect in their moral competence and therefore their moral character. It may surprise us that their concern for these irrelevant features is vicious, but it should not astonish us. After all, their moral concern is fundamentally *misdirected*; given the preceding discussion, it should be clear that misdirected moral concern cannot constitute good will. And given our reasons for adopting the more expansive conception of moral competence defended in this chapter, we should be willing to accept that this misdirected concern represents a kind of moral vice. It is unsurprising that we feel much less inclination to condemn agents like Susan than agents like George; since moral concern for insects is unlikely to result in bad actions, we are are not used to thinking of it as

vicious.

The implication that these agents are also *blameworthy* is more surprising, and potentially more problematic. For my part, I am willing to accept it. Setting aside positive moral incompetence for the moment, it seems that there are sometimes cases of agents who are blameworthy for acting harmlessly and even for acting rightly. I may choose to save a drowning child because it will make my ex-partner jealous, which we may suppose is a morally bad end. Saving the child is still the right thing to do, but I express a desire for the actual bad when I do it. It seems in this case that we should accept the implication that I am blameworthy. If we are willing to accept that agents can be blameworthy for right actions that result from negative moral incompetence, there seems to be no reason why we should not accept that agents can be blameworthy for right (or morally neutral) actions that result from positive moral incompetence.

But suppose that we are unwilling to accept this implication. If so, we can avoid it fairly easily by way of a minor modification. As it stands, MW attributes blameworthiness whenever an attitude satisfying one of conditions (a) through (e) is expressed in an action. On the modified version in question, MW would attribute blameworthiness whenever an attitude of the relevant kind was expressed by an action *that is wrong or bad*. Simply put, we can require that an action actually be bad in order for an agent to be blameworthy for performing it. This will exclude the implication that Susan is blameworthy for deciding not to be a farmer, as well as the implication that Elaine is partially blameworthy for failing to torture the prisoner.[96] As noted, I do not feel compelled to adopt this modification, and prefer to accept the implication that agents like Susan really are blameworthy; for the remainder, I assume that an action does not need to be wrong or

96 It will also imply that I am not blameworthy when I save a drowning child to spite my ex-partner. If it seems to us as though I *would* be blameworthy in this case, then this implication is a cost of the modification.

bad for an agent to be blameworthy for it.

IV. Implications and Conclusion

Because psychopaths systematically fail to respond to the considerations that are actually morally important, we can understand them as agents who display negative moral incompetence on a massive scale. Given the preceding discussion of moral competence, it should be possible for us to imagine a hypothetical counterpart to the psychopath who displays massive *positive* moral incompetence; call this agent the *inverse psychopath*. Whereas the regular psychopath fails to respond to a wide range of features that are actually morally important, the inverse psychopath is morally responsive to an enormous assortment of morally *irrelevant* features – he responds to one such feature for every feature that the regular psychopath *fails* to respond to, we may suppose. The inverse psychopath might feel morally obligated to avoid stepping on cracks in the sidewalk, to refrain from casting shadows on daffodils, and to make every purchase with exact change. And while it seems pretheoretically obvious that there would be something unusual and problematic about such an agent, the expansive account of moral competence defended here implies that the inverse psychopath is deeply morally *vicious*. The inverse psychopath, by stipulation, is comparably morally incompetent to the regular psychopath, and thus is comparably deficient in moral competence – which, on AG accounts, is equivalent to moral virtue.

The inverse psychopath is an imaginary character.[97] Nevertheless, it seems likely that

97 That is, it seems unlikely to me that there actually are any inverse psychopaths, and I do not here assert that any such agents exist. Interestingly, if it turned out that a moral error theory were correct, then it *would be* the case that there actually are inverse psychopaths. Since the truth of an error theory would imply that *no* features are morally significant, it would be the case that morally normal individuals, who respond morally to a wide range of features, are massively positively incompetent. This implication may seem odd – if an error theory is correct, then normal agents are extremely morally vicious! But I think that this is an implication which we should accept – if an error theory were true, then a morally perfect agent would be one who was morally responsive to nothing. If this possibility strikes us as unacceptable, then this should be treated as a consideration against moral error theories, rather than against my claims about moral incompetence.

there are many "localized" instances of positive moral incompetence in the real world. Clinic Bomber Plus was intended to be more-or-less realistic, and, as noted, there are many other real-world cases that are likely to share the same structure. Ideologically-motivated terrorists generally believe that they are acting rightly, and their actions may reflect responsiveness to features which are actually morally irrelevant. War criminals and various other architects of atrocities may act out of sincere moral concern for morally unimportant considerations – they may be attempting to promote ethnic homogeneity, for instance, because they sincerely believe it to be morally good. Many real-world wrongdoers of this kind may display negative incompetence as well – the reasoning that produced their beliefs may reflect insufficient concern for others, or their stated motivations may be a cover for sadistic or self-interested motivations. But I suspect that at least of some of them will be like George, in that their actions will *not* reflect a lack of concern for anything that is actually morally important. If so, we will need to appeal to positive moral incompetence to explain how they can be blameworthy.

Suppose, however, that I am wrong about this, and that agents like George do not occur in real life. Even so, the fact that positive moral incompetence is vicious will still have some practical consequences for the assessment of blameworthiness. For even if there are no agents who are *solely* positively incompetent, there will be many agents who are positively incompetent *in addition* to being negatively incompetent. That is, even if there are no clinic bombers who have the requisite level of moral concern for women, there will be many who have unwarranted moral concern for fetuses *in addition* to their lack of concern for women. And although we will not need positive moral incompetence to explain the fact *that* these agents are blameworthy, we may need it to understand the full *extent* of their blameworthiness – the addition of positive incompetence makes them more vicious, and, plausibly, more blameworthy than they would

otherwise have been.

Thus far in this dissertation, I have not discussed the *degree* to which agents are vicious in virtue of having certain attitudes. This is a significant question that would require an extensive discussion of its own, and considerations of space preclude the possibility of a full treatment here. But the preceding discussion of positive incompetence offers the opportunity to propose a first pass at an account of degrees of viciousness; an interesting implication will be that the vice represented by positive incompetence can sometimes be quite severe. Plausibly, it seems that proper moral responsiveness to the features of actions will result in a set of attitudes with varying strengths. I ought to form an anti-attitude towards stealing, as stealing is a wrong-making feature. But I ought to form a *stronger* anti-attitude towards killing, presumably because killing is morally worse than stealing. And it seems that an agent who entirely lacked anti-attitudes towards killing would be worse than one who entirely lacked anti-attitudes towards stealing. Presumably, this is because the absence of an attitude towards killing represents a greater deviation from moral competence than does the absence of an attitude towards stealing; an agent who is missing an attitude that should be strong is more morally defective than an agent who is missing an attitude that should be relatively weak.

Applied to positive moral incompetence, this seems to imply that agents are more vicious when they form strong attitudes towards irrelevant considerations than when they form weak ones. An irrelevant feature warrants an attitude of zero strength. An attitude with *any* strength represents a deviation from moral competence, but a stronger attitude represents a greater deviation than a weaker one. This is a significant result, because many real-world agents have *very strong* attitudes towards features that may turn out to be morally irrelevant. Some agents, for instance, apparently believe that terminating a fetus is morally equivalent to murder, and form

correspondingly strong anti-attitudes towards it.[98] If it turns out that the well-being of fetuses is morally irrelevant, then these agents will turn out to be extremely vicious – they will have deviated from the ideal of moral competence as dramatically as an agent who has no anti-attitudes towards actual murder, and thus would seem to be comparably vicious.

The fact that different features of actions seem to warrant pro- and anti-attitudes with different strengths raises an interesting possibility: Perhaps we should consider an even more expansive conception of moral competence that requires an agent to respond to each relevant feature with an appropriate attitude of the appropriate *strength*. It does seem as though agents can be vicious by possessing attitudes that are too strong or too weak. An agent might have an anti-attitude towards murder that is relatively weak and easily overridden; most of us would describe this agent as morally vicious. Or an agent might care about something morally important to too great a degree – she might be so averse to lying that she refuses to do so in any circumstances. Once again, this agent seems to be vicious. Perhaps we should understand these cases as examples of negative and positive incompetence, respectively. On this more expansive conception, negative incompetence might consist either in a lack of responsiveness or in insufficiently strong responsiveness to relevant considerations; positive incompetence might consist either in responsiveness to irrelevant considerations or in excessively strong responsiveness to relevant ones. And perhaps the criteria for blameworthiness could be extended as well, so that agents are also blameworthy when their actions display insufficient or excessive responsiveness to relevant considerations.

While this even more expansive conception of moral competence is appealing, I decline

98 A casual Internet search for “abortion” and “murder” will reveal that a fair number of agents feel quite strongly about the impermissibility of this practice.

to endorse it here. This conception faces at least one significant problem, in that it apparently cannot account adequately for agents who are praiseworthy for supererogatory actions. Agents who perform such actions often do so because they have a greater degree of moral concern than is required, and such actions are generally taken to be evidence of virtue. Taken at face value, however, this even more expansive account of moral competence would seem to imply that such agents suffer from a defect in moral character – after all, they apparently care more than they should about certain features of their actions.[99] Perhaps this problem concerning supererogation can be worked out, in which case this even broader conception of moral competence is likely to be viable.[100] For the remainder of this dissertation, I set this question aside and return, finally, to the problem posed by cases of moral disagreement.

I argued previously that many agents who mistakenly believe certain practices to be *permissible* are morally vicious – they show a lack of responsiveness to the features that make these practices wrong, and thereby display negative moral incompetence. I left open the question of how we should evaluate agents who mistakenly believe certain practices to be *impermissible*. Given the discussion in this chapter, we are now in a position to conclude that many of *these* agents will also be vicious. In responding to morally irrelevant features as though they provide moral reasons, these agents display positive moral incompetence. Insofar as their positive moral incompetence is reflected in their actions, these agents are also blameworthy.

---

99 Note that my account of positive moral incompetence does not have this implication; in order to be positively incompetent on my account, an agent must care about something that is entirely morally *irrelevant*; presumably the stereotypically virtuous agents who perform supererogatory actions care about morally *important* considerations to an unusual degree.

100 One way this problem could be resolved is if a given feature warrants an attitude not with a particular strength, but rather with a *range* of strengths, all of which would be acceptable. We could then understand those agents who perform supererogatory actions as those whose attitudes towards right-making features fall towards the "high" end of the acceptable strength range. This schema would allow for the possibility of agents whose attitudes are so strong that they fall outside the acceptable range altogether; such agents, presumably, would be positively incompetent. See Massoud (2016) for an account of supererogation that may be in this ballpark.

There is reason to think that we should revise our judgments of many real-world agents in light of these observations. As noted in Chapter Five, there are many controversial moral practices, and disagreements over these practices can be widespread. *Someone* has to be wrong about each such practice; and, if the mistake is due to faulty responsiveness to the action's features, rather than to false beliefs about which features are present, the party that is in the wrong will be morally vicious. Because there are so many controversies of this kind, it seems likely that each of us will turn out to be wrong about at least *one* of them. It also seems likely that many of us will act in ways that *reflect* our vicious attitudes. An agent need not bomb an abortion clinic in order to express his vicious concern for fetuses, nor perform or procure an abortion in order to express his vicious indifference towards them. Agents can display these attitudes in a variety of less dramatic ways – by voting for certain candidates, by expressing their opinions out loud, or by allowing their moral views to subtly influence their interactions with others. At a minimum, it seems that millions of agents are likely to be morally worse – and more blameworthy for their actions – than we might previously have believed.

Although the main implication of this dissertation may seem to be a gloomy one, it also serves to underscore the value of moral philosophy. Arpaly presents the 2003 version of her AG account in a book entitled *Unprincipled Virtue*, and a major goal of AG accounts is to make room for the possibility of such virtue – virtue that is "naturally-occurring," and which agents like Huck Finn can spontaneously express without the aid of moral theorizing. This is an important goal, as there *is* such a thing as naturally-occurring virtue, and agents *can* be virtuous, at least to a degree, without moral training. But, as the discussion in this dissertation illustrates, there is a limit to how far naturally-occurring virtue can take us. Some moral problems are genuinely hard, and part of virtue consists in responding in the right way to features the moral significance of

which may not be obvious. Arpaly is right to point out that virtue consists in caring about the right things, and that it does not, in principle, require true moral beliefs. In many real-world cases, however, we may require true moral beliefs in order to know *which* things we ought to care about. In the end, naturally-occurring virtue may not adequately equip an agent to navigate the difficult moral questions encountered in modern life. But if we come to better understand which considerations are morally important, we can, hopefully, adjust our attitudes accordingly. Thus, the best way of becoming more virtuous is likely to be through moral philosophy.

References

Aristotle. (1998). *Nicomachean Ethics*. Trans. David Ross. *Oxford University Press*.

Arpaly, N. (2002). "Moral Worth." *Journal of Philosophy* 99(5):223-245.

Arpaly, N. (2003). *Unprincipled Virtue*. Oxford University Press.

Arpaly, N. (2006). *Merit, Meaning, and Human Bondage*. Princeton University Press.

Arpaly, N. and Schroeder, T. (2014a). *In Praise of Desire*. Oxford University Press.

Arpaly, N. and Schroeder, T. (2014b). "Replies to Critics." *Philosophy and Phenomenological Research* 89(2):509-515.

Arpaly, N. and Schroeder, T. (2016). "Response to Swanton and Badhwar." *Journal of Value Inquiry* 50:445-448.

Benn, P. (2000.) "Freedom, Resentment, and the Psychopath." In *Philosophy, Psychiatry, and Psychopathy: Personal identity in mental disorder.* Ed. Christopher Heginbotham. Ashgate Publishing. pp.29-45.

Bradley, G. and George, R. (1995). "Marriage and the Liberal Imagination." *The Georgetown Law Journal* 84:301-320.

Brock, S. (2012). "The Puzzle of Imaginative Failure." *The Philosophical Quarterly* 62(248):443-463.

Carruthers, P. (2011). *The Opacity of Mind*. Oxford University Press.

Chalmers, D. (1996). *The Conscious Mind.* Oxford University Press.

Cleckley, H. (1964.) *The Mask of Sanity.* Fourth Edition. The C.V. Mosby Company.

Driver, J. (2008). "Imaginative Resistance and Psychological Necessity." *Social Philosophy and Policy* 25(1):301-313.

Elster, J. (2011). "How Outlandish Can Imaginary Cases Be?" *Journal of Applied Philosophy* 28(3):241-258.

Eshleman, A. (2014). "Moral Responsibility." In *The Stanford Encyclopedia of Philosophy.* Ed. Edward N. Zalta. http://plato.stanford.edu/entries/moral-responsibility/. Accessed 18 July, 2016.

Fine, K. (2012). “Guide to Ground.” In *Metaphysical Grounding*. Ed. Fabrice Correia and Benjamin Schnieder. Cambridge University Press. pp.37-80.

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

Greenspan, P. (2003). “Responsible Psychopaths.” *Philosophical Psychology* 16(3):417-429.

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

Haji, I. (2010). “Psychopathy, Ethical Perception, and Moral Culpability.” *Neuroethics* 3:135-150.

Hare, R. (1993.) *Without Conscience: The Disturbing World of the Psychopaths Among Us.* The Guilford Press.

Harman, E. (2011). “Does Moral Ignorance Exculpate?” *Ratio* 24(4):443-468.

Hume, D. (2000). *A Treatise of Human Nature*. Ed. David Faye Norton and Mary J. Norton. *Oxford University Press*.

Hurka, T. (2014). “Many Faces of Virtue.” *Philosophy and Phenomenological Research* 89(2):496-503.

Huxley, A. (1932.) *Brave New World*. Chatto & Windus.

Gendler, T.S. (2000). “The Puzzle of Imaginative Resistance.” *Journal of Philosophy* 97(2):55-81.

Gendler, T.S. and Liao, S. (2016). “The Problem of Imaginative Resistance.” In *The Routledge Companion to Philosophy of Literature*. Ed. John Gibson and Noël Carroll. Routledge. pp.405-418.

Kerstein, S. (2013). *How to Treat Persons*. Oxford University Press.

Kiehl, K.A. (2008). “Without Morals: The Cognitive Neuroscience of Criminal Psychopaths.” In *Moral Psychology, Volume 3.* Ed. Walter Sinnott-Armstrong. The MIT Press. pp.119-149.

Leary, S. (Forthcoming). “Non-Naturalism and Normative Necessities.” *Oxford Studies in Metaethics* 12.

Levy, N. (2007). “The Responsibility of the Psychopath Revisited.” *Philosophy, Psychiatry and Psychology* 14(2):129-138.

Levy, N. (2015). *Consciousness and Moral Responsibility.* Oxford University Press.

Liao, S., Strohminger, N., and Sripada, C.S. (2014). "Empirically Investigating Imaginative Resistance." *British Journal of Aesthetics* 54(3):339-355.

Libet, B. (1985). "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action." *The Behavioral and Brain Sciences* 8:529-566.

Litton, P. (2008). "Responsibility Status of the Psychopath: On Moral Reasoning and Rational Self-Governance." *Rutgers Law Journal* 39:349-392.

Lyon, D.R. and Ogloff, J.R.P. (2000). "Legal and Ethical Issues in Psychopathy Assessment." In *The Clinical and Forensic Assessment of Psychopathy.* Ed. Carl B. Gacono. Lawrence Erlbaum Associates. pp.139-173.

Maibom, H. (2005). "Moral Unreason: The Case of Psychopathy." *Mind and Language* 20(2): 237-257.

Maibom, H. (2008). "The Mad, the Bad, and the Psychopath." *Neuroethics* 1:167-184.

Markovits, J. (2010). "Acting for the Right Reasons." *Philosophical Review,* Vol. 119, No. 2: 201-242.

Markovits, J. (2012). "Saints, Heroes, Sages, and Villains." *Philosophical Studies* 158: 289-311.

Massoud, A. (2016). "Moral Worth and Supererogation." *Ethics* 126:690-710.

Model Penal Code. (1981). *The American Law Institute*. §4.01(2).

Nelkin, D. (2015). "Psychopaths, Incorrigible Racists, and the Faces of Responsibility." *Ethics* 125(2):357-390.

Pope Paul VI. (1968). *Humanae Vitae*. Papal Encyclical.

Plato. (1961). *Euthyphro*. Trans. Lane Cooper. In *The Collected Dialogues*. Ed. Edith Hamilton and Huntington Cairns. Princeton University Press. pp.169-185.

Robbins, P. (2009). "Modularity of Mind." In *The Stanford Encyclopedia of Philosophy.* Ed. Edward N. Zalta. http://plato.stanford.edu/entries/modularity-mind/. Accessed 14 July, 2016.

Rosen, G. (2004). "Skepticism About Moral Responsibility." *Philosophical Perspectives* 18(1):295-313.

Rosen, G. (2010). “Metaphysical Dependence: Grounding and Reduction.” In *Modality: Metaphysics, Logic, and Epistemology.* Ed. Bob Hale and Aviv Hoffmann. Oxford University Press. pp.109-136.

Sartorio, C. (2011). “Actuality and Responsibility.” *Mind* 120(480):1071-1097.

Scanlon, T.M. (2008). *Moral Dimensions*. Belknap Press of the Harvard University Press.

Schaffer, J. (2009). “On What Grounds What.” In *Metametaphysics: New Essays on the Foundations of Ontology.* Ed. David J. Chalmers, David Manley, and Ryan Wasserman. Oxford University Press.

Scott, R. (2014). “Psychopathy – An Evolving and Controversial Construct.” *Psychiatry, Psychology, and Law* 21(5):687-715.

Shoemaker, D. (2011). “Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility.” *Ethics* 121(3):602-632.

Shoemaker, D. (2015.) *Responsibility from the Margins.* Oxford University Press.

Star, D. (2011). “Two Levels of Moral Thinking.” *Oxford Studies in Normative Ethics* 1:75-96.

Star, D. (2015). *Knowing Better: Virtue, Deliberation, and Normative Ethics*. Oxford University Press.

Strawson, P. (1962). “Freedom and Resentment.” *Proceedings of the British Academy* 48:1-25.

Stear, N. (2015). “Imaginative and Fictionality Failure: A Normative Approach.” *Philosophers' Imprint* 15(34).

Talbert, M. (2008). “Blame and Responsiveness to Moral Reasons: Are Psychopaths Blameworthy?” *Pacific Philosophical Quarterly* 89(4):516-535.

Talbert, M. (2012). “Accountability, Aliens, and Psychopaths: A Reply to Shoemaker.” *Ethics* 122(3):562-574.

Talbert, M. (2014). “The Significance of Psychopathic Wrongdoing.” In *Being Amoral: Psychopathy and Cognitive Incapacity.* Ed. Thomas Schramme. The MIT Press. pp.275-300.

Terzo, S. (2013). “Life Begins at Conception, Science Teaches.” *Live Action News*. http://liveactionnews.org/life-begins-at-conception-science-teaches/

Väyrynen, P. (2013). “Grounding and Normative Explanation.” *Aristotelian Society Supplementary Volume* 87(1):155-178.

Walton, K. (1994). “Morals in Fiction and Fictional Morality.” *Proceedings of the Aristotelian Society* 68:27-50.

Weatherson, B. (2004). “Morality, Fiction, and Possibility.” *Philosophers' Imprint* 4(3).

Wegner, D. (2002). *The Illusion of Conscious Will*. The MIT Press.

Yablo, S. (2002.) “Coulda, Woulda, Shoulda.” In *Conceivability and Possibility.* Ed. Tamar S. Gendler and John Hawthorne. Oxford University Press. pp.441-492.

Sean Clancy
541 Hall of Languages, Syracuse, NY 13244 | (443)-244-2930 | sclancy@syr.edu

Areas of Specialization: Normative Ethics, Moral Responsibility
Areas of Competence: Free Will, Philosophy of Mind, Applied Ethics (esp. Bioethics)

Education

Syracuse University
PhD, Philosophy (dissertation defended 10/21/2016) 2011-2017

University of Maryland
MA, Philosophy 2009-2011

Princeton University
BA, Philosophy (*summa cum laude*) 2005-2009

Publications

"Psychopaths, Ill-Will, and the Wrong-Making Features of Actions" Forthcoming
Ergo

"A Strong Compatibilist Account of Settling" 2013
Inquiry 56 (6):653-665

Active Revise-and-Resubmits

"Imaginary Moral Reasons and Positive Moral Incompetence"
Ethics

Reviews

Review of Neil Levy's *Consciousness and Moral Responsibility* 2015
(With Travis Timmerman)
The Philosophers' Magazine

Presentations

"Psychopaths, Ill-Will, and the Wrong-Making Features of Actions" Forthcoming
Seminar on Human Knowledge and Action
East China Normal University, Shanghai, China

"Psychopaths, Ill-Will, and the Wrong-Making Features of Actions" Forthcoming
Mary Hatch Marshall Award Ceremony
Syracuse University, Syracuse, New York

"Psychopathy, Responsibility, and Normative Explanation" 2016
Gothenburg Responsibility Conference
University of Gothenburg, Gothenburg, Sweden

"Transhuman Lives and the Critical Level of Well-Being" 2016
Interdisciplinary Workshop on Human Enhancement
University of Tübingen, Tübingen, Germany

"Imaginary Moral Reasons and Positive Moral Incompetence" 2015
International Ethics Conference: Reasons and Virtues
Australian Catholic University, Melbourne, Australia

"Blameworthiness and Imaginary Moral Reasons" 2015
American Philosophical Association Central Meeting
St. Louis, Missouri

"Moral Ignorance and Imaginative Resistance" 2014
Free Will and Moral Responsibility Summer School
Moscow Center for Consciousness Studies, Moscow State University

"A Strong Compatibilist Account of Settling" 2013
Workshop on *A Metaphysics for Freedom*
Center for the Study of Mind and Nature, University of Oslo

Papers in Progress

"Imaginative Resistance as a Methodological Hazard"

"Posthuman Lives and Critical-Level Views of Well-Being"

"Altruism, Desert, and Organ Markets"

Awards

Mary Hatch Marshall Award for Graduate Research 2017
College of Arts and Sciences, Syracuse University

Outstanding Teaching Award 2015
The Graduate School, Syracuse University

Summer Research Grant 2015
Philosophy GSO, Syracuse University

Certificate of University Teaching – Future Professoriate Program 2015
The Graduate School, Syracuse University

Alexander Guthrie McCosh Senior Thesis Prize 2009
Department of Philosophy, Princeton University

John Martyn Warbeke Senior Thesis Prize 2009
Department of Philosophy, Princeton University

Dickinson Senior Thesis Prize 2009
Department of Philosophy, Princeton University

Teaching Experience

Syracuse University:

As Instructor of Record:

PHI 109: Introduction to Philosophy (Honors)	Fall 2016
PHI 107: Theories of Knowledge and Reality	Fall 2016
PHI 393: Contemporary Ethics	Spring 2015
PHI 192: Introduction to Moral Theory	Fall 2014
PHI 107: Theories of Knowledge and Reality	Spring 2014
PHI 107: Theories of Knowledge and Reality	Fall 2013

As Teaching Assistant:

PHI 293: Ethics and the Media Professions	Spring 2017
PHI 197: Human Nature	Spring 2015
PHI 293: Ethics and the Media Professions	Fall 2015
PHI 293: Ethics and the Media Professions	Spring 2013
PHI 191: Ethics and Contemporary Issues	Fall 2012
PHI 107: Theories of Knowledge and Reality	Spring 2012
PHI 192: Introduction to Moral Theory	Fall 2011

University of Maryland:
As Teaching Assistant:

PHIL 282: Free Will and Determinism	Spring 2011
PHIL 100: Introduction to Philosophy	Fall 2010
PHIL 100: Introduction to Philosophy	Spring 2010

Southside Charter School:
As Visiting Instructor:

"Philosophy"	Fall 2014,
Co-instructor for philosophy class for 8th grade students	Fall 2015

Professional Service

Anonymous peer review — 2016
Erkenntnis

Graduate Teaching Mentoring Program — 2016
Syracuse University Department of Philosophy

Graduate Conference External Speaker Co-Coordinator — 2015
Syracuse University Graduate Conference

Graduate Conference Committee — 2013
Syracuse University Graduate Conference

Dissertation

What Counts as Desiring the Actual Good? (defended 10/21/2016)

On *actual good* (AG) accounts, virtue consists in having the appropriate attitudes towards those considerations that are identified as good and bad by the correct normative theory; agents are praiseworthy for those actions that reflect the appropriate attitudes, and blameworthy for those that reflect inappropriate attitudes. However, I argue that existing AG accounts suffer from a defect that prevents them from providing unambiguous evaluations of moral character and moral worth in a range of interesting and realistic cases. To remedy this defect requires us to fundamentally reexamine the relationship between the correct normative theory and the actual good. I argue that rather than picking out a limited number of discrete goods and bads, a normative theory identifies a range of features that make actions right or wrong by performing normative work at various levels. For an AG account to be successfully applied to the full range of interesting cases, we will require a further, substantive story about which attitudes, towards which kinds of right- and wrong-making features, *count* as "desires for the actual good" – or the *appropriate* attitudes, for the purposes of assessing moral character and moral worth. After offering such a story, I apply my newly strengthened AG account to previously problematic cases, such as those of the psychopath and of the agent with unusual moral beliefs.

Committee:
Ben Bradley
Hille Paakkunainen
David Sobel
Mark Heller
Nomy Arpaly (Brown University)