BETWEEN VIRTUE AND VICE:
MORAL WORTH FOR THE REST OF US

by

WILLIAM MATHIEU DOUCET

A thesis submitted to the Department of Philosophy
In conformity with the requirements for
the degree of Doctor of Philosophy

Queen's University
Kingston, Ontario, Canada
August, 2009

Abstract

Most of us fall short of virtue—we are, at various times, weak-willed, selfish, self-absorbed, hypocritical, morally complacent, cowardly, and self-deceived. But most of us are not vicious, either. In this dissertation I argue that the actions of flawed, morally imperfect agents can be as praiseworthy as the actions of more perfectly virtuous people.

The first, introductory chapter explains my account of moral worth, which depends on the assessment of an agent's deliberative outlook in acting. In the second chapter, I argue that being praiseworthy on every possible occasion is not a precondition for being praiseworthy on any particular occasion. This may seem obvious, but it is also inconsistent with a common interpretation of the nature of virtue.

The third chapter argues that someone's actions can be morally worthy despite displaying a failure of practical rationality quite similar to weakness of will, or akrasia. By exploring cases of so-called inverse akrasia, I argue that sometimes, an agent can be praised for acting in ways that he himself believes are morally wrong, and that while these actions display serious failures of practical reason, they can still be both done for a good reasons and deserving of praise.

The fourth chapter explores the moral status of hypocrisy. I reject the standard interpretation of hypocrites as blameworthy manipulative deceivers, and argue instead that they are people who misdirect their ethical attention by caring too much about their image for having certain values, and not enough about the values themselves.

The second, third, and fourth chapters draw a close connection between moral imperfection and failures of self-knowledge. The fifth and final chapter therefore considers the nature of such failures of self-knowledge by exploring the moral

significance of self-deception. I argue that, in a central range of cases, it is impossible to be self-deceived about the content of one's own mind. Instead, I argue that the morally relevant form of self-deception is a failure of self-assessment. This has important implications for our understanding of moral development, since it means that such development centrally involves the cultivation of a specific kind of self-knowledge.

Acknowledgments

This project is, in part, an appeal for leniency for those of us with flaws and imperfections, brought about by the conviction that this describes all of us. Naturally, this project itself contains many flaws and imperfections, but it certainly contains far fewer than it would have without the excellent supervision of Rahul Kumar, whose enthusiasm, support, and friendly criticism made my work immeasurably better and helped me see this project through to completion. I also owe a great deal to Stephen Leighton, who taught me all that I know about Aristotle and whose questions have always helped me to make my thinking clearer.

This project was made possible by the generous support of the Department of Philosophy at Queen's University, and particularly Sergio Sismondo, David Bakhurst, Deborah Knight, Judy Vanhooser, and Marilyn Lavoie.

I would also like to acknowledge the Social Sciences and Humanities Research Council and the Ontario Graduate Scholarship for the financial support that enabled me to complete this dissertation.

Finally, thanks to my parents, Joan and Alfred, for their unflagging support, and most of all to Megan and Rosa, for support and joy in equal measure.

Table of Contents

Abstract i
Acknowledgments iii
Table of Contents iv
Chapter 1: Introduction 1
Chapter 2: Occasional Virtue 22
Chapter 3: In Praise of Akrasia? 61
Chapter 4: What Is Wrong With Hypocrisy? 96
Chapter 5: When Self-Knowledge Fails 136
Bibliography 168

Chapter 1: Introduction

Most of us fall short of virtue—we are, at various times, weak-willed, selfish, self-absorbed, hypocritical, morally complacent, cowardly, and self-deceived. We do things we shouldn't, and we don't do the things we should. But most of us are not vicious, either. There may be a big difference between those who are genuinely virtuous and the rest of us, but that difference should not be overstated: those of us who are not virtuous are not all vicious, amoral egoists. True, some people really *are* vicious, but most people who lack virtue do not spend all of their time breaking solemn promises, mugging little old ladies, and betraying those who love them. In fact, if virtue is as rare as most philosophical accounts of virtue make it out to be, then very few of us are fully virtuous, and so most people who lack virtue generally *act* in a thoroughly permissible manner at least most of the time. Non-virtuous people keep their promises, pay their taxes, obey the law, and care for their loved ones.

Not only can such flawed, imperfect people act as morality requires: we often praise them for doing so, believing that they deserve moral credit for their actions. They may not be praiseworthy all that often, and they certainly do not act in morally worthy ways as reliably as virtuous, saintly people, but it appears that even the most weak-willed, selfish, and morally complacent people sometimes find it within themselves to do the right thing, and to do so in ways that deserves praise.

But perhaps such appearances are deceptive. Maybe, despite appearances to the contrary, praise is not something available to imperfect, non-virtuous agents, even when

they act in accordance with morality's demands. After all, even when it is clear that someone has acted just as morality requires, we often hesitate to praise him, perhaps because we believe that he was not properly motivated. People can keep their promises because they are afraid not to or because they care about social conformity rather than morality, they can obey the law simply to avoid going to jail, and they can care for sick family members in the hopes of an inheritance. Such people might act in conformity with the demands of morality, but they hardly merit praise for doing so.

Perhaps this view seems unduly harsh, but there are several influential strands in philosophical moral psychology that suggest that imperfect, non-virtuous agents are never deserving of praise. Those who defend a moral psychology that gives an important place to the concept of virtue frequently argue—and sometimes simply assume—that it is not possible for someone to be virtuous merely *some* of the time. To have the virtue of courage, for example, is to be *characteristically* courageous, to act courageously whenever courage is required, and so it is argued that no one who lacks the virtue of courage can ever act courageously.[1]

[1] For those who hold that virtue is unified—that is, to have any one virtue is to have all of them—examples multiply dramatically. On this view, someone who seems brave but intemperate not only lacks the virtue of temperance, he *also* lacks the virtue of bravery. Two different classical version of the unity thesis can be found in Plato, *Protagoras*, trans. C.C.W. Taylor (Oxford: Oxford University Press, 1996), and Aristotle, *Nicomachean Ethics*, trans. Terence Irwin (2nd edition; Indianapolis: Hackett, 1999). Modern defences of the same thesis can be found in, for example, Terry Penner, 'The Unity of Virtue', *The Philosophical Review*, 82 (1973), 35-68, and Neera Badhwar, 'The Limited Unity of Virtue', *Noûs*, 30 (1996), 306-32. An excellent recent critical discussion of the thesis is Susan Wolf, 'Moral Psychology and the Unity of the Virtues', *Ratio,* XX (2007), 145-67.

This does not mean, of course, that imperfect people never keep their promises, help others, love their friends, or face danger for a worthy cause. Non-virtuous agents might, on occasion, *appear* to act justly, or generously, or courageously. Nevertheless, as they no not have the relevant virtue, there is some important difference between their actions and those of the genuinely virtuous agent. This difference means that, while they might act in morally desirably ways, they do not deserve any praise for doing so. Praise, on this view, is restricted to the virtuous, and closed off to those who are flawed, imperfect, and otherwise less than fully virtuous.

Kant, in fact, worries that we may be systematically mistaken in praising *anyone*, regardless of how reliably they do the right thing. He is "willing to admit out of love for humanity that most of our actions are in accordance with duty",[2] but this does not mean that any of those actions are morally worthy and deserving of praise. In fact, he admits that it is possible that no one has ever actually acted from the motive of duty. After all, our motives are often obscure, not only to others, but even to *ourselves*: perhaps self-interest and self-love lie behind actions we believe have the purest of motives. As Kant observes, "one need not be exactly an enemy of virtue, but only a cool observer who does not take the liveliest wish to be straight off its realization, in order to be doubtful at times whether any true virtue is to be found in the world."[3]

There are, then, powerful strands within moral psychology that suggest that the actions of flawed, imperfect agents rarely or never merit praise. My aim in this dissertation is to show that this is mistaken—in particular, my aim is to show that flawed,

[2] Immanuel Kant, *Groundwork for the Metaphysics of Morals*, trans. James W Ellington (3rd edition; Indianapolis: Hackett, 1993) at 19.
[3] Ibid. p. 20.

imperfect, and otherwise less than virtuous agents can act in ways that are not simply morally *desirable*, but morally *worthy*. This project therefore explores the conditions of moral worth. I argue that flawed, imperfect agents—which is to say, most if not all of us—can not only act in accordance with the requirements of morality; when they do so, they can merit praise in just the same way as more saintly and virtuous agents. I even show how it is possible for flawed agents to merit praise at the same time as they display their flaws: not only can akratic agents and hypocrites overcome their imperfections to act in just the same praiseworthy way as virtuous agents, there are even cases in which they can merit praise for actions which are, at the very same time, akratic or hypocritical. Since most of us are less than fully virtuous without being close to wicked, these are among the most important issues moral psychology can address, since they have to do, not with idealized and theoretical agents, but you and me.

In this introduction, I explain how I understand the distinction between an action's moral desirability and its moral worth, and outline the conditions that must be satisfied in order for an agent's actions to be praiseworthy. This account of moral worth serves as the foundation for the chapters that follow.

I. Moral Rightness and Moral Worth

We sometimes hesitate to praise people who, it seems clear, have done the right thing. It is common for extremely wealthy people to donate large sums of money to charitable organizations, universities, hospitals, and other worthy causes, and these donations are often greeted with scepticism or even cynicism. Sceptics point out that the sums of money involved might seem large to us, but that they are the equivalent of pocket change

to the very wealthy, and there is nothing particularly noble about giving away something you will not much miss. There is also a deeper source of scepticism about such donations: it is not simply that the sacrifices involved are small, but that they are not really *sacrifices* at all, since the donations are motivated by self-interest, rather than generosity. The suspicion is that the benefactors are primarily interested in earning tax breaks, or in feeding their egos by getting their names on buildings, or in getting good public relations, and are not, in the main, overly concerned with the moral worth of the causes they are supporting. Any tax break, or building, or PR boon would do, and it just so happens that morally worthy causes fit the bill. A cynic might agree that the Bill and Melinda Gates Foundation does an enormous amount of good in the world, and that those who carry out its works or donate money to it do morally good things, and yet deny that Bill and Melinda Gates themselves ought to be praised for funding the Foundation.

Surely such complete cynicism about the motives of wealthy benefactors—or anyone else—is misplaced. Nevertheless, a degree of scepticism is often warranted: rich and poor alike often do good things for reasons that are unconnected to whatever makes their actions good. The hesitation we feel in praising those whose actions we recognize as morally good illustrates that there is a distinction between, on the one hand, whether or not an action is morally *right*, and on the other, whether or not the agent deserves any *praise* for her action. I will call this the distinction between an action's moral *desirability* (its rightness or wrongness) and its moral *worth* (the extent to which it merits praise or

blame).[4] T.M. Scanlon picks out much the same idea when he distinguishes an action's "permissibility" from its "meaning."[5]

The distinction between an action's desirability (or permissibility) and its worth (or meaning) depends on the reasons, motives, or intentions of the agent who performs it. We can accept that donating a million dollars to charity is a good thing and yet be hesitant about the moral worth of the action, because we can be suspicious of the benefactor's motives: was he aiming to help the needy, or to earn a tax break? As Scanlon explains the distinction, whether an action is permissible (and so, in some contexts, whether it is good, right, or obligatory as well) does not generally depend on the agent's intentions. What *does* depend on the agent's intentions is the action's meaning: "the significance, for the agent and others, of that agent's willingness to perform that action for the reason he or she does."[6] When we speak of someone doing "the right thing for the wrong reason," we are picking out this distinction, and recognizing that the moral worth of an agent's actions—the extent to which he is praiseworthy for acting in the way that he does—is distinct from the rightness of that action, since the former, but not the latter, depends on the reasons for which he acted.[7]

---

[4] I borrow this terminology from Nomy Arpaly, *Unprincipled Virtue* (Oxford: Oxford University Press, 2003) at 69. As Arpaly points out, this deviates somewhat from the Kantian use of "worth," since an action can have both positive moral worth (can be *praiseworthy*), or it can have negative moral worth (can be *blameworthy*).

[5] T.M. Scanlon, *Moral Dimensions: Permissibility, Meaning, Blame* (Cambridge, MA: Harvard University Press, 2008).

[6] Ibid. at 3.

[7] "What makes an action wrong is the consideration or considerations that count decisively against it, not the agent's failure to give these considerations the proper weight," and the same holds, *mutatis mutandis,* for what makes an action right. Ibid. at 23.

An action's desirability and its moral worth are obviously closely related: whether or not someone deserves praise or blame for her actions will depend in large part on whether what she did was right or wrong, desirable or undesirable. Right actions are good candidates for praise, wrong actions good candidates for blame. But the connection between these is not perfect. Someone can do something morally right and yet not deserve praise, as the example of the person who gives money to a charitable foundation for the tax receipt shows. And we can even stretch the imagination to consider cases where it would be appropriate to *blame* someone for acting in ways that are morally desirable. J.S. Mill considers such a case, imagining a tyrant who rescues a man from drowning only so that he can torture him later. We can think it a good thing that the man was rescued and still blame the tyrant on the grounds that his motives were vicious.[8]

The same holds true, *mutatis mutandis,* for the connection between wrongness and blame. Not only can people, such as Mill's tyrant, be blamed for acting in permissible ways, but some people who act in ways that are clearly impermissible do not deserve blame for doing so. They may be excusably ignorant of the wrongness of their actions, or they may have acted in distress, or under compulsion, in ways that mitigate the extent to which blame is appropriate.[9]

The reason we can distinguish between an action's moral desirability and its moral worth is, in part, because a morally right action can be performed for reasons that

[8] John Stuart Mill, *Utilitarianism* (Indianapolis: Hackett, 2001) at 18 n2. Mill denies that the tyrant's action is the same as the person who rescues him for the right reason, since the tyrant only rescues him as part of an extended act "far more atrocious." As Scanlon points out, that still does not mean that the tyrant should leave the man to drown. Scanlon, *Moral Dimensions: Permissibility, Meaning, Blame* at 219n3.

[9] A point emphasized in P.F. Strawson, 'Freedom and Resentment', in Gary Watson (ed.), *Free Will* (Oxford: Oxford University Press, 1982).

are wholly unconnected to the reasons that make the action right. Moral worth is an evaluation of not merely the action, but of the agent's reasons for carrying it out. A person who donates money to charity in order to collect a tax break acts for reasons that have nothing to do with the rightness of his donation. He wants to advance his own economic self-interest, and it just so happens that this leads him to act in a morally desirable way. This is not simply a happy accident, since the tax break was likely designed to encourage people to donate to worthy causes, but any praise surely rests on the authors of the tax policy, not the individuals who take advantage of it to save money and thereby happen to do something morally desirable.

If, on the other hand, someone donates money only so that people will think that he is a good person, there does seem to be *some* connection between the reasons for which he donates and the reasons that make donating right, but the connection nevertheless seems to be the wrong one. The *right* reason to donate, and the reason that would make his action morally worthy, is that his donation can contribute in some small way to easing suffering and rectifying injustice: the *wrong* reason is that he recognizes that other people believe this, and so will think that he is a good person if he donates.[10]

The distinction between an action's rightness and its moral worth, then, depends on the distinction between an action that is merely in *conformity* with the demands of morality and an action that is in *compliance* with those demands: it is possible to do the right action for merely self-interested reasons, rather than because it is the in fact the right

[10] I will say more about such misdirected ethical attention in Chapter 4, where I discuss the nature of hypocrisy.

action.[11] This distinction is a familiar one, and plays an important role in otherwise competing account of moral worth. It is reflected, for example, in Kant's argument that there is no moral worth in the actions of the shopkeeper who, out of self-interest, never cheats his customers. His action "certainly accords with duty… but this is not nearly enough for making us believe that the merchant has acted this way from duty and the principles of honesty."[12]. The shopkeeper actions are morally desirable, and conform to morality's demands: it is a good thing that he treats his customers honestly. Nevertheless, says Kant, his actions have no moral worth. We might be happy that the shopkeeper is honest, but his actions only "fortunately light" upon what duty requires. While he acts in a way that is consistent with what morality requires, he does not *comply* with morality, since he does not act not act honestly *because* that is what morality requires.

Aristotle draws attention to a similar distinction by contrasting states that are "merely… in accord with the correct reason" with those "involving correct reason."[13] Only the latter, says Aristotle, are virtuous. In his discussion of bravery, he points out that there are other things that look like bravery: those who defend the city in battle because they are afraid of their superiors, or because they are paid to do so, or because they are ignorant of the dangers, appear to us to be brave. But these people are not brave, since they do not act for the right reason which, for Aristotle, is for the sake of the fine.[14] Mill, too, recognizes that the same action can be done for many different motives, and though

[11] The language of conformity and compliance clearly laid out in Chapter 1 of Joseph Raz, *Practical Reason and Norms* (2nd edition; Oxford: Oxford University Press, 1990).
[12] Kant, *Groundwork for the Metaphysics of Morals* at 10.
[13] Aristotle, *Nicomachean Ethics* at 1144b28.
[14] Ibid. Book III, chapter 8.

he claims that what an agent's motives are "has nothing to do with the morality of the action", he happily admits that it is relevant to "the worth of the agent."[15]

Aristotle, Kant, and Mill disagree strongly about which actions are right, what makes those actions right, what makes right actions morally worthy, and even the extent to which we are capable of acting in a praiseworthy fashion. They all agree, however, that an action that is praiseworthy does not merely *conform* with the demands of morality, but *complies* with those demands. A praiseworthy action, in other words, is one that is done for the right reasons.

If this is the right account of moral worth, then to claim that the actions of flawed, imperfect, non-virtuous agents are never morally worthy even when they are morally desirable is to claim that such imperfect agents can never act for the right reasons. In fact, this account of moral worth will be qualified below: while the reasons for which an agent acts are central to the moral worth of his actions, other, related considerations play a role as well. Nevertheless, I will argue that the actions of non-virtuous agents can be praiseworthy, and that they can even merit praise when those actions display the sorts of imperfections that explain the agents' lack of virtue.

II. The Deliberative Outlook View of Moral Worth

So far, the discussion of moral worth has made it seem as an agent is praiseworthy for doing the right thing just in case he is did it for the right reason. In Arpaly's words, "what

[15] Mill, *Utilitarianism* at 18.

we praise in the morally praiseworthy agent is her responsiveness to moral reasons."[16] This obviously captures an important aspect of moral worth. But is it a complete explanation of the conditions of praise and blame? It seems to me that two people can act for precisely the same reason and yet deserve praise to different degrees. This is because the reasons for which someone acted is only one (albeit centrally important) element on the deliberative outlook he has in acting, and it is this deliberative outlook, and not merely the reason, that ultimately matters to our assessments of moral worth.

In speaking of an agent's deliberative outlook on his action, I have in mind a combination of factors. These include the reason for which he acted, of course, but it can also include much more: the relationship that reason stands in to other potential reasons and actions, his emotional orientation to the reason, and the force with which that reason moves him. Different people can do the same action for the same reasons while *seeing* those reasons in very different ways: one person might see the reason as obviously and decisively correct, while another might deliberate long and hard about whether the reasons are good ones, and so act hesitantly, uncertain that she is doing the right thing. It is possible for one person to embrace the reasons and welcome the actions they call for, and for another to acknowledge that the same reasons obtain, but see the actions they require as involving a significant sacrifice, and so act reluctantly. The possibility of

[16] Arpaly, *Unprincipled Virtue* at 72. Arpaly points out that what matters is not that the agent *believe* that those reasons are moral, but that the reasons are, *in fact*, moral reasons. Believing that one's reasons are moral ones is not sufficient for one's actions to merit praise: we are all too aware that people can falsely believe that morality requires them to do the most heinous things. And believing that one's reasons are moral ones is not necessary, either. In the discussion of inverse akrasia in Chapter 3, I offer an argument for the potential praiseworthiness of agents who falsely believe that their reasons are morally wrong. In the discussion of hypocrisy in Chapter 4, I explore agents who falsely believe their reasons are morally right.

akrasia even shows that it is possible for someone to act for reasons he judges to be bad ones. The same consideration can be the decisive reason for which several agents act, and yet take many different places within those agents' deliberative outlooks.

In speaking of this as a deliberative *outlook*, I do not intend to suggest that it need be reflexive or conscious. It can be—indeed, it generally is—something about which the agent is quite unaware. We can even be *mistaken* about our deliberative outlooks: we can be mistaken about the reasons for which we act, or about how strongly those reasons move us, or about which competing desires and considerations might play a role in our view of the situation. Indeed, such failures of self-knowledge will play an important role in my explanations of the various morally significant failures of practical reasoning that I discuss in this work. Nevertheless, a deliberative outlook is the way an agent is deliberatively oriented towards a situation; it is not generally (though it can be) the way he is oriented towards *himself*.

Finally, in speaking of an agent's deliberative outlook, I have in mind the particular orientation an agent takes toward a specific situation at a given time. This should be distinguished from an agent's characteristic way of looking at similar situations at different times, and from her general appreciation of the force of her reasons across a wide range of situations. The idea of a person's outlook sometimes carries this second meaning as well: when we say someone has a 'gloomy outlook on life', for example, we mean that he is characteristically pessimistic and inclined to grumble. I prefer to reserve 'deliberative outlook' for an agent's orientation to a specific case, and so distinguish between an agent's deliberative outlook and her deliberative *dispositions*. The former is a snapshot of her orientation towards her reasons for action on a particular occasion; the

latter is a global or characteristic tendency to, for example, be motivated by certain considerations and unmoved by others.

Quite clearly, an agent's deliberative dispositions influence her particular deliberative outlooks in many ways, not least by partially constituting the range of possible outlooks. To have the deliberative dispositions of virtue is to be reliably inclined to adopt the deliberative outlook of virtue, case by case, and of course there may well be a range of reasons that can only play a role in the deliberative outlooks of agents with particular dispositions. Nevertheless, in Chapter 2 I will argue that agents who do not have fully virtuous deliberative *dispositions* can, on some occasions, adopt fully praiseworthy deliberative *outlooks*.

The moral significance of an agent's deliberative outlook is at the heart of neo-Aristotelian accounts of virtue. In moral contexts, akratic, self-controlled, and virtuous people can all recognize the same reasons they have to act morally, and they can all judge that they ought to act on the basis of those reasons, and they can all form the same intention to act on those reasons. We all know, however, that wanting to do the right thing, knowing what the right thing is, and even fully intending to do the right thing, is not always sufficient for us to actually *do* it. We are sometimes weak and irresolute: despite our best intentions, we can fail to do the right thing. Akratic agents know how they ought to act, but they act otherwise; they form the right judgements, but they give in to the temptation. Self-controlled and virtuous agents, on the other hand, both manage to act as they intended, on the basis of the reason that all three agents recognized. But there is still a difference between the self-controlled and the virtuous person, which is revealed by a similarity between the akratic and the self-controlled agent. Both of them are

tempted to do something other than what morality requires, though the self-controlled person manages to overcome his temptation.

The virtuous person, however, stands apart from both the akratic and the self-controlled person, since she alone *is not tempted.* She willingly accepts her duty, and discharges it with pleasure. She does not see the demands of morality as a burden she has to bear and which she is tempted to abandon; rather, she see it as an essential component of her happiness.[17] Both the self-controlled and the virtuous person do the same action, and for the same reason, but they nevertheless have *different* deliberative outlooks on their actions: one of them welcomes the action, while the other does it reluctantly and while being tempted to act otherwise; one of them sees the reason as decisive, while the other weighs it against competing considerations. The virtuous person is therefore more praiseworthy than the self-controlled person. Both have acted in the same way, and for the same reason, but they have done so while displaying *different* deliberative outlooks, and this difference explains the greater praise that the virtuous person deserves.

[17] "Pleasure" and "happiness" need to be properly interpreted in this context. Aristotle draws a close connection between virtue, pleasure, and happiness: he claims "virtue is about pleasure and pains" (1104b16) and one of the themes of the *Nicomachean Ethics* is that only the virtuous person can be truly happy. But this does not mean that all virtuous actions are pleasant: while "the end that bravery aims at seems to be pleasant", that does not mean that the brave person *enjoys* risking his life. In fact, there would be something worrying about someone who did. The brave person will endure the risk of death, however, "because that is fine or because failure is shameful." So the active exercise of every virtue need not be pleasant: "it is pleasant only insofar as we attain the end." Aristotle, *Nicomachean Ethics* at 1117b1-20. Nevertheless given that his city is at risk, there is nothing the brave person would rather be doing, even if he does not take *pleasure* in risking his life. Second, most virtuous actions are not like risking one's life to defend one's city. The virtuous person *does* enjoy being kind to others and doing favours for friends. He sees these as components of a good life, rather than as burdens that duty forces him to bear.

This is not to say that the self-controlled person deserves *no* praise: Aristotle himself seems to allow that "continence and resistance seem to be good and praiseworthy conditions."[18] But because the virtuous person is someone whose desires are shaped by correct reason, while the self-controlled person's are not, the Aristotelian view is that the person with the virtue of temperance is *more* praiseworthy than the person who is merely self-controlled. Even if an action is praiseworthy in virtue of being done for the right reason, the *degree* of praise depends on the agent's deliberative outlook, not simply his reason for acting.

This emphasis on the importance of an overall deliberative outlook, over and above an agent's reason for action, is not specifically Aristotelian. Nomy Arpaly, whose account of moral worth is not Aristotelian, also argues that two actions done for the identical reason can merit different degrees of praise. She agrees that to be praiseworthy is to have acted for the relevant moral reasons, but claims, further, that "an agent is more praiseworthy, other things being equal, the deeper the moral concern that has led to her action."[19] To have a concern for morality is to have a desire to follow its commands: the action of someone whose desire is stronger is more praiseworthy than the action of someone whose desire is weaker, even if both people did the same action for the same reason. The strength of an agent's desire to follow morality is therefore a component of her deliberative outlook on actions in moral contexts, and so Arpaly, too, sees an agent's

[18] Ibid. at 1145b9. "Continence" here is equivalent to "self-control." In saying that such people "seem" to be praiseworthy, Aristotle leaves open the possibility that this belief is mistaken. His subsequent discussion, however, and particularly VII.9, suggests that he in fact agrees, as he there describes continence as "excellent" and a mean between excess and deficiency. (1151b25-30) When he compares continence to temperance, however, it is clear that the latter is more praiseworthy than the former.

[19] Arpaly, *Unprincipled Virtue* at 84.

deliberative outlook, and not simply the reason for which he acted, as determining the extent to which his actions are praiseworthy.

Arpaly's account is in some respects similar to Harry Frankfurt's, for whom the assessment of someone's actions (and her character) depends in part on, not simply the desires that moved her to act, but also on whether she identifies with those desires and endorses them.[20] Frankfurt's view of what a deliberative outlook is differs sharply from Aristotle's—in fact, Frankfurt would likely object to the overly cognitive tone of the idea of a *deliberative* outlook, and prefer to speak of an agent's *volitional* outlook or his volitional structure in acting.[21] But all of the views mentioned above agree on a centrally important truth: there is more to the moral worth of an action than the reasons for which it was done. An agent's moral worth depends in large measure on the attitudes with which she considers those reasons, be it pleasure, regret, dismay, alienation, or endorsement. Her emotional response to the reasons, the force with which they move her, and the way she sees them as relating to other, potentially reason-giving considerations all play a role in determining the moral worth of his actions, and so it is possible for two people to do the same action for the same reason and yet deserve very different degrees of praise or blame.

While they both acknowledge this important truth about the importance of moral worth, however, both Arpaly's and the Aristotelian account imply, in different ways, either that the actions of imperfect, less than fully virtuous agents are never praiseworthy,

[20] This account of caring is developed throughout the papers collected in Harry Frankfurt, *The Importance of What We Care About* (Cambridge: Cambridge University Press, 1988).

[21] One of Frankfurt's themes is that "the essence of being a person lies not in reason but in will." 'Freedom of the Will and the Concept of a Person', ibid., p 17.

or that, to the extent that they are, they are never as praiseworthy as the actions of the fully virtuous person. I believe that this is a mistake. While it is certainly true that flawed, imperfect agents are less reliably praiseworthy than the perfectly virtuous, and that they generally display less concern for morality than their more saintly neighbours, this does not mean that none of their actions can be fully praiseworthy, or that they can deserve, at most, only an attenuated degree of praise. In my view, not only can imperfect agents act in ways that are fully praiseworthy, but they can even merit a level of praise for actions that display the same imperfections for which they are otherwise justifiably blamed. Praise is not reserved for a select, virtuous few: it is accessible to the flawed, imperfect many as well.

III. Summary of the Dissertation

My aim in this work is, in part, modest. I simply want to show that the actions of imperfect, flawed, less than fully virtuous agents can be morally worthy. Perhaps this seems so obvious that it hardly bears saying, let alone dedicating an entire dissertation to developing arguments in its favour. In a sense, I agree. I do think that it is obvious that flawed and imperfect people can act in morally worthy ways, because I think that almost all of us are flawed and imperfect in important respects, and yet I think that moral worth is not an impossible standard to meet. An adequate account of moral worth, in other words, should make it at least possible that, in many cases, when we praise someone's actions, we are *warranted* in doing so. We can of course makes mistakes, and praise

those who do not deserve it.[22] But in my view, any account of moral worth that is out of the grasp of most if not all of us is an account that we should reject.

I believe that it is important that our understanding of moral worth be adequate to actual moral agents because I conceive of any convincing account of moral worth as part of a larger account of moral psychology. Moral psychology, in turn, has two aims: it is a descriptive enterprise, aiming to explain how actual moral agents act, deliberate, form intentions, and feel, and it is a normative enterprise, aiming to assess the extent to which agents do these things well or badly. A satisfactory moral psychology should combine these two aims, and so its normative ideals should be constrained by its descriptive content. This means that any account of moral worth should be informed by an understanding of moral agents as they actually are and can reasonably be expected to be. In order to assess whether someone acted for the right reasons, whether his cognitive, emotional, and motivational outlook on his actions deserve praise, blame, or indifference, it is important that we understand what is involved in actual agents acting for a reason, and what sorts of overall deliberative outlooks are available to them. Praise and blame are assessments of real agents, not idealized ones, and so they should be based in standards that real agents can meet. That is not to say that the standards of praise and blame should be lax or permissive; they may be extremely demanding, so long as the demands are not

[22] The question of whether praise is *warranted*, or whether someone is *worthy* of praise, is slightly different form the question of whether it would be a good idea to actually praise him. Praising someone who does not deserve it can be a way of encouraging him morally desirable behaviour, perhaps in the hopes that he will eventually acquire the right deliberative outlook. This is how the moral education of children often proceeds. And we should sometimes withhold actually praising someone for doing something praiseworthy—perhaps it would embarrass him, or cause others to resent him. See Arpaly, *Unprincipled Virtue* at 70-1.

misplaced, and are understood to apply to moral agents as we find them in the world. My aim is to show that imperfect agents can meet these standards, however demanding.

In the second chapter, I argue that being praiseworthy on every possible occasion is not a precondition for being praiseworthy on any particular occasion. Someone can, for example, be praised for acting courageously even if they are not always courageous, and so the actions of flawed and imperfect people can, on many occasions, have precisely the same moral worth as the actions of the perfectly virtuous. This may seem obvious, but it is also inconsistent with a common interpretation of the nature of virtue. This leads me to re-examine the way we carve up the distinction between virtue and its absence, and to argue that full virtue is an aspirational ideal, rather than a state of character one either possesses or lacks.

The second chapter, then, deals with cases where the relationship between praise and imperfection is *diachronic:* agents who are praiseworthy at some times and imperfect at others. But the argument I make leaves open the possibility that, in order to merit praise, the person has to have acted perfectly. All it requires is that the person does not *always* act perfectly. The third and fourth chapters, however, deal with more troubling cases, where the relationship between praise and imperfection is *synchronic:* agents who are praiseworthy while, at the very same time, displaying an important failing or imperfection.

The third chapter argues that someone's actions can be morally worthy despite displaying a failure of practical rationality quite similar to weakness of will, or akrasia. By exploring cases of so-called inverse akrasia, I argue that sometimes, an agent can be praised for acting in ways that he himself believes are morally wrong, and that while

these actions display serious failures of practical reason, they can still be both done for a good reason and morally worthy, deserving of praise.

The fourth chapter explores the moral status of hypocrisy. Hypocrisy is widely condemned as involving self-interested pretence to virtue: not only are hypocrites never praised, but even when they act in ways that comply with the demands of morality, they are often *blamed* for engaging in self-serving deception. I reject this overly harsh view of hypocrites, and offer a competing view of hypocrisy. On my view, hypocrites are not blameworthy, manipulative deceivers; rather, they are people who misdirect their ethical attention by caring too much about their image for having certain values, and not enough about the values themselves. But this does not mean that hypocrites care nothing for such values. Hypocrites do not care about their values *enough:* this does not mean that they do not care about them *at all.* I therefore argue that, while hypocrisy is certainly blameworthy, hypocrites can be praised for acting in ways that do genuinely reflect their values.

The second, third, and fourth chapters discuss different ways in which imperfect agents can merit praise, and so they necessarily discuss the various ways in which moral agents fall short of virtue. Each chapter emphasizes a slightly different failure of practical reason, but all three failures—unreliability, akrasia, and hypocrisy—involves a clear failure of self-knowledge. The fifth chapter therefore considers the nature of such failures of self-knowledge by exploring the moral significance of self-deception. I argue that the failures of self-knowledge that beset the self-deceived—and so, by extension, imperfect agents generally—are not failures to know the content of their own minds. In fact, I argue that, in a central range of cases, it is impossible to be self-deceived about the content of

one's own mind. Instead, I argue that the morally relevant form of self-deception is a failure of self-assessment. This has important implications for our understanding of moral development, since it means that such development centrally involves the cultivation of a specific kind of self-knowledge.

Chapter 2: Occasional Virtue

We are complicated and hard to pin down: even the most apparently virtuous among us occasionally falls prey to weakness or temptation, and seemingly vicious and nasty characters sometimes surprise us with what appears to be generosity or courage. Many of us demonstrate a deep commitment to some moral causes, and yet are completely unmoved by, or even blind to, relevantly similar causes. It can be hard to know how to assess such characters: should we praise them for their occasional successes and blame them for their failures, or do those failures show us that their virtues are only illusory?

Consider, for example, the character of Lord Darlington, from Kazuo Ishiguro's novel *The Remains of the Day*. Lord Darlington has obvious moral flaws: through the course of the novel he consorts with fascists of various stripes, fires his Jewish employees, and ends his life with a reputation as a notorious Nazi-sympathizer.

Despite his flaws, Lord Darlington himself was no vicious fascist. Even his critics acknowledged that he was a fine, generous, and honourable man. He served in the First World War, but bears no ill-will towards the German people. As he said: "I fought that war to preserve justice in this world… I wasn't taking part in a vendetta against the German race."[1] He was therefore deeply distressed by the suffering he witnessed on visits to Germany in the early 1920s. He blamed this suffering on the conditions of the Treaty of Versailles, conditions he believed to be unjust: in his view, "once you've got a

[1] Kazuo Ishiguro, *The Remains of the Day* (New York: Vintage International, 1988) at 73.

man on the canvas, that ought to be the end of it. You don't then proceed to kick him."[2] Moved by the suffering he witnessed, and by his sense that it was the result of an unjust and dishonourable treaty, Lord Darlington hosted several diplomatic meetings aimed at renegotiating the terms of that treaty. It is through these seemingly honourable efforts that he eventually came to consort with Nazis.

Stevens, Darlington's butler and the novel's narrator, is aware (though perhaps not as aware as he should be) of Lord Darlington's faults. But Stevens remains convinced that Lord Darlington was a good man. Darlington's efforts may have been tragically misplaced, and he may, have been horribly naïve and utterly blind to the injustices of the Nazi regime, but Stevens believes that he was, nevertheless, a just and honourable man whose efforts deserve the highest praise. But is Stevens right? Should we praise Darlington's initial efforts and motives, even as we recognize that he was naïve, complacent, and occasionally thoughtless, and that his sense of justice and honour was eventually twisted to serve truly evil ends? Or does his subsequent failure to recognize obvious injustice reveal to us that he was never truly moved by a praiseworthy sense of justice at all, and that something else motivated him from the start?

Fiction, and no doubt real life, is full of such morally complicated characters, who seem to mix honour with cruelty, virtue with vice. But how should we assess them? When they act in ways that seem morally desirable, do they deserve praise? To suppose that these characters—from Lord Darlington to the prodigal son to the loveable rogue to the honourable thief—deserve praise is to suppose that they manage (though perhaps only on rare occasions) to adopt a morally worthy deliberative outlook on their actions. But

[2] Ibid. at 87.

perhaps we are mistaken in attributing such an outlook to them it may be that such an outlook is one that they simply cannot adopt. How could Lord Darlington ever be moved by an admirable sense of justice, if he so clearly lacks a fully-fledged commitment to justice, and indeed seems to be a promoter of the most horrendous injustice? In other words, how could he adopt a praiseworthy deliberative outlook if he lacks a praiseworthy deliberative disposition? Accounts of moral worth that emphasize the importance of the virtues draw a very close connection between an agent's deliberative dispositions and his particular deliberative outlooks, and so such accounts often seem to imply that morally worthy actions are the exclusive province of the virtuous: non-virtuous people may often act in morally desirable ways, but something about their lack of virtue means that, their actions are not, at the same time, praiseworthy.

In this chapter, I come to the defence of flawed and imperfect agents: sometimes, such people can deserve praise for their actions. I therefore take issue with a central element of the standard view of virtue, which makes the moral worth of actions derivative on the possession of a virtuous character. As a result, I argue that the gap between virtuous and non-virtuous agents is not as wide as this standard view claims.

I. The standard view of virtue

Denying that Lord Darlington was motivated by a sense of justice and honour in first concerning himself with foreign affairs, *merely* on the grounds that this sense of justice was not fully present in all aspects of his life has an objectionable whiff of moralism about it. If he has acted precisely as the just person would have acted, what possible

reason could we have to deny that he has acted justly, besides an unjustifiable distaste for those who are less than perfect?

But that is not the issue. The claim that Lord Darlington never acted justly or honourably is not simply a restatement of the claim that he does not have the virtue of justice. Rather, the claim is that, because he does not have the virtue of justice, he *cannot* act precisely as the just person does, and so his actions cannot be morally worthy. This does not require the wholly implausible argument that non-virtuous agents can never even act in mere conformity with the demands of morality. What it does require is denying that non-virtuous agents can do so in the appropriate, praiseworthy way by arguing that there is necessarily some difference between the deliberative outlooks each of them takes on their actions in moral contexts. If there is indeed such a difference, and if an action's moral worth depends on the deliberative outlook of the agent who performed it, then the actions of non-virtuous characters like Lord Darlington can never be as morally worthy as the actions of properly virtuous agents. Praising such imperfect agents is therefore always unwarranted.

Denying praise to Lord Darlington might jar with our intuitions, or at least the intuitions of filmmakers and novelists, but it seems to be implied by what I will call "the standard account of virtue." This account, which is at the heart of virtue theoretic approaches to ethics and which is prominent in contemporary moral psychology, is broadly Aristotelian (which is not to say that it is necessarily Aristotle's). It explains right action, and so morally worthy action, in terms of the virtues, rather than defining virtues simply in terms of a statistical tendency to do right actions. As John McDowell puts it,

this view approaches the question of 'how should one live?' "*via* the notion of a virtuous person. A conception of right conduct is grasped, as it were, from the inside out."[3]

A virtue, on this view, is a "firm and unchanging state," which is to say a characteristic disposition to behave reliably in certain ways and to take certain considerations as reasons for action.[4] Such a disposition "issues in nothing but right conduct."[5] A virtuous person is therefore someone who consistently, reliably, does the right thing, and with the right deliberative outlook, case-by-case. In fact, for the versions of the standard view endorse the thesis that virtue is unified and to have any virtue is to have all of them, any agent who fails in the domain of any virtue thereby fails in all of them: anyone who is intemperate is thereby unkind, and one who is less than courageous is thereby unjust.

The standard view is therefore what Tom Hurka calls a "dispositional view." It begins with the notion of the virtuous disposition, and it then defines particular praiseworthy actions and deliberative outlooks "derivatively, as ones that proceed from such dispositions."[6] The upshot of such a view seems to be that is that only those with

---

[3] John McDowell, 'Virtue and Reason', in Roger Crisp and Michael Slote (eds.), *Virtue Ethics* (Oxford: Oxford University Press, 1997), 141-62 at 141. Similarly, Rosalind Hursthouse defines right action as follows: "an action is right iff it is what a virtuous agent would characteristically… do in the circumstances," where this is not a truism, but rather a substantive account of right action that requires filling in the concept of a virtuous person. Rosalind Hursthouse, *On Virtue Ethics* (Oxford: Oxford University Press, 1999).

[4] Aristotle, *Nicomachean Ethics* at 1105b1.

[5] McDowell, 'Virtue and Reason', at 143.

[6] Thomas Hurka, 'Virtuous Act, Virtuous Dispositions', *Analysis*, 66 (2006), 69-76 at 70. Aristotle puts the point as follows: "for actions in accord with the virtues to be done temperately or justly it does not suffice that they themselves have the right qualities. Rather, the agent must also be in the right state when he does them." Aristotle, *Nicomachean Ethics* at 1105a29-23.

virtuous dispositions can act in praiseworthy ways, since being praiseworthy is derivative on having the firm and unchanging disposition of virtue. As Aristotle says, "we are praised or blamed… insofar as we have virtues or vices."[7]

On this view, then, those who do *not* do the right thing reliably, case-by-case, are not virtuous. Since, by definition, non-virtuous people *lack* the firm and unchanging disposition of virtue, they clearly cannot act from virtue on those occasions when their actions are morally desirable: an unchanging disposition is not something that we can adopt only some of the time. This view therefore draws a sharp contrast between virtuous and non-virtuous agents. Those who have the firm and stable disposition and who always do the right thing are virtuous: those who do not have the disposition and who are do not reliably do the right thing are not. For the sake of argument, I will also speak of 'virtuous' and 'non-virtuous' agents, though in the end I will conclude that the standard view draws this distinction too sharply.

My aim in this chapter is not to reject wholesale the insights of the standard account of virtue. Indeed, I share with the defenders of the standard account Anscombe's view that it is only profitable to do moral philosophy if we are equipped with an adequate moral psychology.[8] I also share many of the central commitments of the account, including the view that an agent's deliberative outlook is central to the moral evaluation of his actions, that certain praiseworthy deliberative outlooks are only available to those with particular deliberative dispositions, and that the best way to reconcile the apparent

[7] Aristotle, *Nicomachean Ethics* at 1106a2-3.
[8] G.E.M. Anscombe, 'Modern Moral Philosophy', *Collected Philosophical Papers Volume Iii: Ethics, Religion, Politics* (Minneapolis: University of Minnesota Press, 1981).

conflict between morality and rational self-interest is by understanding the nature of the virtuous person's distinctive practical outlook. In arguing for these positions, defenders of the standard view have, quite understandably, relied on idealized accounts of what is involved in virtue. This has left them open to a series of empirically influenced objections, most prominently from defenders of so-called situationist moral psychology, which argues, on the basis of research in empirical psychology, that no one in fact possesses the sorts of robust and stable character traits that the standard account of virtue describes.[9] My aim, in a sense, is to defend the standard account from such attacks, by showing that its central insights do not merely apply to the idealized agents that populate the pages of philosophical books and articles, but also apply to the sorts of flawed, imperfect agents that populate the world. This will of course involve criticizing the idealizations of the standard account, but largely in the service of showing that such idealizations are not necessary.

I therefore share with the defenders of the standard account the view that it is an agent's deliberative outlook that determines whether his morally desirable actions are praiseworthy. My aim is merely to show that a praiseworthy deliberative outlook is not only open to those idealized virtuous agents who always and reliably act in morally desirable ways: I will argue that many flawed and non-virtuous agents can, on occasion, act in ways that are morally worthy. Mediating this dispute requires exploring just what is

[9] See, for example, Gilbert Harman, 'Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error', *Proceedings of the Aristotelian Society,* 99 (1999), 315-31. and John Doris, *Lack of Character: Personality and Moral Behaviour* (Cambridge: Cambridge University Press, 2002). For an argument that situationism does not properly interpret the empirical evidence from which it draws its conclusions, see Gopal Sreenivasan, 'Errors About Errors: Virtue Theory and Trait Attribution', *Mind,* 111 (2002), 47-68.

involved in the standard view's account of the deliberative outlook of virtue in order to determine whether non-virtuous people can ever adopt it. The answer to this question, in turn, will have implications for the plausibility of some elements of the standard account of virtue, since it will cast into serious doubt the plausibility of the distinction between virtuous and non-virtuous agents.

II. Reliability and the counter-factual condition

On the standard view, the virtuous agent is someone who does the right thing case-by-case because she is reliably and characteristically sensitive to the status of certain considerations as compelling reasons for action, and she gives those reasons the proper place within her deliberative outlook. A kind person notices situations in which she could make others happy, takes the fact that she could make someone happy as providing her a good reason for taking the trouble to do so, and, when kindness is appropriate, she takes it as a decisive reason for doing so. The non-virtuous agent, by contrast, is not reliably sensitive to the reason-giving force of such considerations: he does not always notice the needs of others, when he does he does not always see them as providing him with a reason to help, and even when he does see it as providing a reason he sometimes wrongly concludes that the reason is outweighed by competing self-interested considerations. This does not mean, of course, that non-virtuous people *never* see the needs of others as providing them with good or even decisive reasons for action, but only that they do not do so consistently and reliably.

This difference between the virtuous and the non-virtuous person provides the foundation for the standard view's argument that non-virtuous agents can never act with

the deliberative outlook of virtuous agents, and so can never merit praise for their morally desirable actions. Here is John McDowell's explanation of what is involved in the virtuous person's sensitivity to reasons:

> A kind person can be relied on to behave kindly when that is what the situation requires. Moreover, his reliably kind behaviour is not the outcome of a blind, non-rational habit or instinct, like the courageous behaviour—so called only by courtesy—of a lioness defending her cubs. Rather, that the situation requires a certain sort of behaviour is (one way of formulating) his reasons for behaving that way, on each of the relevant occasions. So it must be something of which, on each of the relevant occasions, he is aware. A kind person has a reliable sensitivity to a certain sort of behaviour which situations impose of his behaviour.[10]

This is in some respects similar to Aristotle's claim that "for actions in accord with the virtues to be done temperately or justly it does not suffice that they themselves have the right qualities. Rather, the agent must also be in the right state when he does them". The right state is one that is "firm and unchanging", which is to say, the right state is *virtue*.[11] In both cases, there is a risk that two thoughts are being run together. The first is that someone with the virtue of kindness reliably behaves kindly whenever the situation calls for kindness (and, *mutatis mutandis,* the same goes for courage, justice, benevolence, and so on). On the standard view, this is true by definition.

The second thought is that a person can only be aware of the reasons for being kind if he is *always* aware of them, and so always *behaves* kindly. This is the foundation of any argument that non-virtuous agents cannot adopt the deliberative outlook of virtue, and it is a much stronger claim. It puts the following counter-factual condition on an agent's acting on, or even recognizing, a reason:

---

[10] McDowell, 'Virtue and Reason', at 142.
[11] Aristotle, *Nicomachean Ethics* at 1105a25-05b1.

> *The counter-factual condition on moral worth: In any particular instance, an agent can only be appropriately sensitive to a consideration's status as a reason if he is* always *sensitive to that status*.

According to the counter-factual condition, an agent has only successfully recognized a reason to be kind here and now if he would *also* recognize that same reason in all other situations in which it is likely to be occur. This counter-factual condition is very strong indeed: it rules out the possibility that someone could recognize a reason to be kind of some occasions and fail to recognize it on others. In fact, since the standard account of virtue claims that only those with the virtue of kindness are sensitive to the reasons for acting kindly, and since the standard view *also* requires that the kind person act kindly in all situations requiring kindness, the counter-factual condition is even stronger: it says that an agent has only successfully recognized a reason to be kind here and now if he would also recognize all of the *other* reasons to be kind in all of the other situations in which kindness is required. So the standard view, as expressed by McDowell, even rules out the possibility that someone could recognize *one* class of reasons to be kind on some or even all occasions, but not recognize a different class of reasons to be kind on other occasions. Someone who has the virtue of kindness recognizes both that the efforts of young children deserve praise and encouragement, and that volunteering at the local homeless shelter is a kind and generous thing to do. If only people with the virtue of kindness can recognize the reasons to be kind, then only those who recognized both sets of reasons—those relating to children, and those relating to the homeless—are capable of recognizing *either* set of reasons. If the standard view is right, then it seems as if someone who does not understand that it would be kind to encourage young children (perhaps

because he has little experience with them) must therefore also fail to fully and properly recognize the reasons he has to volunteer at the shelter, even if in fact he does so volunteer. If the standard account's counter-factual condition were correct, then all virtuous or praiseworthy actions would indeed involve a sort of knowledge that is completely inaccessible to the non-virtuous agent, since the only way to be sensitive to *any* reason to be kind would be to be sensitive to *every* such reason.

If this is true, and if an agent can only recognize a reason to be kind, or courageous, or just, if he always recognizes *all* such reasons, then we cannot explain the actions of agents like Lord Darlington by claiming that he was appropriately sensitive to the relevant reasons, since it is clear that Lord Darlington often fails to act on, and perhaps even to recognize, the many reasons he has to be just across a wide range of situations. Just people do not fire their Jewish employees, socialize with fascists, or work to advance the aims of the Nazi regime. One alternative way of explaining a case like Lord Darlington's is suggested by McDowell's example of the lioness defending her cubs—after all, his initial visits to post-war Germany were undertaken in order to visit a friend, and so perhaps it was just visceral disgust at the impoverished condition of his friend, rather than a genuine appreciation of injustice, that moved him to action. But why should the fact that Lord Darlington later shows himself to lack a fully developed sense of justice be sufficient, on its own, for us to conclude that his actions *must* have been the result of something other than the appropriate recognition that punishing German civilians for the sins of their political leaders is cruel, unjust, and the cause of much preventable suffering? In other words, what reason do we have for supposing the counter-factual condition to be true?

III. Objections to the counter-factual condition

There are several reasons for doubting the truth of McDowell's counter-factual condition. First, it seems that someone could be sensitive to the reasons to be kind in many cases where kindness is called for without thereby acting on those reasons. Someone who is not reliably kind might recognize that someone else's need provides him with a good reason for helping, but still not help, because, weighing the reasons he has to be kind against other, competing considerations, he falsely concludes that some other reason is stronger. In this case, though he makes a mistake in failing to act on the reasons he has to be kind, he is not completely *insensitive* to these reasons. So it seems perfectly possible that someone could be reliably sensitive to all of the reasons he has to be kind, and yet not act in a reliably kind way, since he might fail to always give the reasons he has to be kind the proper place in his deliberative outlook. A reliable sensitivity to the existence of reasons to be kind, in other words, need not entail reliably kind behaviour, even in someone who is genuinely motivated to act for those reasons. So the actions of someone who is not reliably kind could, on this view, result from a perfectly reliable sensitivity to the existence of reasons to be kind. If agents are only praiseworthy if the reasons for which they act are the deliverances of a reliable sensitivity, then the kind actions of a certain sort of non-virtuous agent seem to be perfectly plausible candidates for praise.

In fact, McDowell is well aware that it is possible for an agent to recognize that a reason obtains without thereby acting on it when he should. But this does not mean that the standard view allows that such agents can be praiseworthy, since being *appropriately* sensitive to a reason involves more than simply knowing that the reason obtains and has

some unspecified motivational force. Central to the virtuous person's characteristic deliberative dispositions is a proper understanding of the ways in which her reasons for action relate to other, potentially reason giving considerations. In specific deliberative outlooks, this often involves properly weighing reasons against one another, but it can involve much more, as well: it is characteristic of the virtuous person to see certain considerations, "not as outweighing or overriding any reasons for acting in other ways which would normally be constituted by other aspects of the situation… but as *silencing* them."[12] This distinguishes the virtuous person from the merely self-controlled person: the virtuous person sees the reasons he has to be kind, or courageous, as decisive, and as completely silencing the reason-giving force of potentially competing considerations. The self-controlled person might act for the same reasons, but still have a very different deliberative outlook on those reasons, since he weighs them against the (perhaps self-interested) considerations he had to act otherwise. The difference between virtue and self-control, says McDowell, "becomes intelligible if we stop assuming that the virtuous person's judgment is a balancing of reasons for and against."[13]

For the virtuous person, then, there can be no such thing as a legitimate reason not to be generous, just, or courageous, and so she does hesitate to do the right thing. She may regret that the circumstances require her to act—the courageous person wishes the city were not in danger, and the just person would prefer the world to be free from injustice—but given that the circumstances do obtain, she does not see the actions they demand of her as involving a sacrifice of her own self-interest. There is nothing she

[12] McDowell, 'Virtue and Reason', at 146. Emphasis added.
[13] Ibid.

would rather do than what virtue requires, and no reason that she recognizes to act otherwise. The merely self-controlled person, by contrast, does acknowledge such reasons, and so he can both hesitate and feel that doing the right action involves a regrettable sacrifice of self-interest. When he does, act, it may be only half-heartedly, as he still feels the pull of the competing self-interested reasons to act otherwise.

To say that, for the virtuous person, certain potentially reason-giving considerations are 'silenced' is not to say that he is completely *ignorant* of their potential reason-giving force. Sometimes, this is true: Bernard Williams has argued in several places that the virtue is sometimes best exemplified by an absence of conscious deliberation, and that to even consider the potential reason-giving force of certain considerations is to merit some degree of blame.[14] But this is not always the case: very often, to say that a consideration is 'silenced' is to say that the agent in question has quite consciously decided to set it aside and exclude from his deliberations. As T.M. Scanlon puts it, being a good teacher, or committee member, or guide, often involves "bracketing the reason-giving force of your own interests which might otherwise be quite relevant and legitimate reasons for acting in one way rather than another."[15] This bracketing can be quite conscious: often, what is important is not whether a potential reason-giving consideration *occurs* to someone, but whether he or she actually treat it as reason-giving

[14] For example, he argues that someone who does the right thing for the right reason can be criticized for having "one thought too many" Bernard Williams, 'Persons, Character and Morality', *Moral Luck* (Cambridge: Cambridge University Press, 1981), 1-19 at 19, and that some moral concerns are best embodied in "deliberative silence." Bernard Williams, *Ethics and the Limits of Philosophy* (Cambridge, MA: Harvard University Press, 1985) at 185.

[15] T.M. Scanlon, *What We Owe to Each Other* (Cambridge, MA: Belknap Press, 1998) at 52.

and allow it to carry deliberative force. So a characteristic feature of the virtuous person's deliberative dispositions is a sensitivity to the status of certain considerations as "exclusionary reasons"—that is to say, second-order reasons to refrain from giving some further consideration any reason-giving force or deliberative weight.[16] Such exclusionary reasons can completely silence what would be, in other situations, perfectly legitimate reasons for action. This suggests a modified version of the counter-factual condition:

> *The counter-factual condition on moral worth (modified): In any particular instance, an agent can only be appropriately sensitive to a consideration's status as a reason if he is* always *sensitive to that status, and this sensitivity involves, not simply recognizing that the consideration has some unspecified reason-giving force, but properly understand that consideration in relation to other potentially reason-giving considerations, where this includes recognizing the consideration's status as an exclusionary reason.*

This modified version of the counter-factual rules out the possibility that someone could reliably recognize a reason and yet not always act on it when he should, since it draws a necessary connection in all cases between being sensitive to a reason and giving it its proper place in one's deliberative outlook. On this view, only those with fully virtuous deliberative dispositions can ever attain fully praiseworthy deliberative outlooks in particular cases.

In addressing the original counter-factual condition, I argued that it is possible for an agent to be reliably sensitive to the reason-giving force of a morally relevant consideration without reliably acting on those reasons. I therefore accepted, for the sake of argument, the central element of the counter-factual condition: the claim that an agent can only recognize the force of a reason if he *always* recognizes that reason. I simply

[16] For a detailed explanation of exclusionary reasons, see Joseph Raz, *Practical Reason and Norms* at 39.

pointed out that we could accept this and still allow that imperfect agents can act on the basis of a reliable sensitivity, and thus merit praise, since such agents can sometimes fail to give each reason its right place in their deliberative outlook when they act. The modified version of the coulter-factual condition, however, rules this out. This forces us to once again address the heart of the counter-factual condition: the claim that one can only be properly sensitive to a reason if one is *always* properly sensitive to that reason. Is this a claim that we have any reason to accept?

IV. Understanding a reason

At the heart of the counter-factual condition, it seems to me, is the conviction that someone who only recognizes the force of a reason on some occasions cannot possibly understand it in quite the same way as the person whose understanding extends across a much wider range of cases. Someone who only keeps promises some of the time—say, only promises made to close friends or promises that do not inconvenience him unduly—simply does not understand what is involved in promise keeping in anything like the way of someone who reliably and consistently keeps his promises. This, in turn, seems implied by a quite plausible view of what is involved in knowing or understating any fact, proposition, or reason.

A child learning basic arithmetic understands the rules of addition when she is consistently able to add small sums and gets the correct answer. If she only gets the right answer some small fraction of the time, we conclude that she does not understand arithmetic, and we rightly suppose that the few times she gets the right answer are due to chance, not to an understanding of arithmetic that comes and goes. To understand the

rules of addition is to be able to apply those rules it a suitable range of cases, and so understanding any particular case will require an understanding of many cases. A child who can only tell us that 2 and 2 are 4 does not understand addition: she has simply memorized a single sum. Likewise, we do not conclude that a child who tells us that 3 and 7 are 10 on one occasion but that they are 8 on anther and 12 on still another understood the sum the first time but the second two; rather, we conclude that the first time was just a lucky guess. Understanding basic addition requires both being able to repeatedly give the correct answer to any particular sum, and being able to give the correct answer to an indefinite range of sums. There is, then, certainly *some* counter-factual component to the understanding, and this includes understanding the force of a reason. Anyone who truly understands a reason in one case understands it in an indefinite range of relevantly similar cases: it is not possible to understand a reason on a single occasion in a single case and yet fail to understand it on all other occasions and in other relevantly similar cases.[17]

Since there is some counter-factual element to understanding a reason, the standard account of virtue can claim that, even in cases where an imperfect agent appears to recognize a the force of a reason and give it its proper place in her deliberative structure, there is *still* something amiss about her understanding of that reason, and that this difference means that the imperfect agent is not entitled to the same level of praise as the virtuous agent. The non-virtuous agent might recognize that her promise to her friend

[17] It is something like this picture of understanding a reason (and a mathematical rule) McDowell defends in 'Non-cognitivism and rule-following', *Wittgenstein: To Follow a Rule* (London: Routledge, 1981). To understanding a reason is to understand how to 'go on in the same way', where this involves, not the application of an algorithm, but the appreciation of something like a form of life.

is an exclusionary reason for her to keep it, silencing all other competing, self-interested reasons not to keep it. But if she does not *always* recognize (in cases where she should) that promises provide her with such exclusionary reasons, then she displays a lack of understanding of the reason-giving force of promises, and this lack of understanding is present even in those cases where she *does* recognize that her promise acts as an exclusionary reason.

This objection acknowledges that the imperfect agent can be sensitive to moral reasons to some extent, but it insists that this is not sufficient to count as the sort of genuine appreciation of the reasons characteristic of virtue, and so it denies that actions done with this deliberative outlook are truly praiseworthy. This view of moral worth is not limited to defenders of the standard account of virtue. Nomy Arpaly, who does not endorse the standard account of virtue, makes a similar claim. For Arpaly, different agents can merit different degrees of praise despite acting for identical reasons: on her view, an agent is "more praiseworthy, other things being equal, the deeper the moral concern that led to her action."[18] Arpaly's explanation of depth of concern is decidedly counter-factual: a person with a deeper moral concern is motivated by, emotionally affected by, and sensitive to moral considerations in situations where the person with a shallower moral concern is not.[19] An action's moral worth depends on the depth of the moral concern of the agent who performs it, and so a person who is motivated by moral concerns in situations where I am not has a deeper concern for morality than I do even if we *both* presently notice, are emotionally affected, and are motivated by the identical

[18] Nomy Arpaly, *Unprincipled Virtue* at 84.
[19] Ibid. at 86-7.

moral reason. His action is therefore more morally worthy than mine, even though both actions are identical. On both the standard view of virtue and Arpaly's depth of concern view, the extent to which an agent's actions deserve praise depends on how that agent would act in a indefinitely large further range of cases, such that two agents could perform the same action, for the same reason, with the same deliberative outlook, and yet merit sharply different degrees of praise.[20] Whether a deliberative outlook can be praised depends on further facts about the agent's deliberative dispositions.

At issue here is the standard account of virtue's insistence that the counter-factual condition be fully satisfied in order for an agent's deliberative outlook on any particular action to be morally worthy. It is certainly true that flawed, non-virtuous agents do not have the same understanding of the reasons they have to be kind, or generous, or courageous, that virtuous people have. They do not understand the full range of application of those reasons, or appreciate all of the ways in which such reasons can silence or exclude other considerations, and they often have an imperfect view of the relative weight those reasons carry in cases where they do not act as exclusionary

[20] Arpaly, it should be said, seems to be somewhat uncomfortable with making this a *necessary* feature of her account, and so she qualifies her explanation of strength of concern in several ways: someone has deeper concern if he is motivated, emotionally affected, and cognitively aware on a wider range of cases, but only "other things being equal." As she points out, in moral life, other things are quite often not equal: if they were moral deliberation would be much easier, and moral agency less fraught with challenges. Moreover, she says that the motivational, emotional, and cognitive dispositions are only "*associated*, other things being equal, with strength of concern." (ibid. p. 85, emphasis in original) So her strength of concern account of moral worth does not simply *define* the praiseworthiness of any particular action in terms of a counter-factual disposition to care in a further range of cases. Nevertheless, her account does draw a very tight connection between the degree of praise an action merits and the agent's overall deliberative orientation towards a further range of cases.

reasons. It is precisely this lack of general understanding that explains their lack of reliability and makes them non-virtuous. But these are all claims about the non-virtuous person's understanding of the reasons' application more generally: they are not claims about their understanding of the application of those reasons in specific cases.

It may well be true that, for someone to understand a reason, he cannot understand it one time only, and he must understand it in an indefinitely large range of cases. The counter-factual condition is legitimate to that extent. But an indefinitely large range of cases should not be mistaken for *all possible* cases. So the claim that understanding a reason involves understanding it in an indefinitely large range of cases in no way entails that he must understand it on *every possible* occasion in order for him to understand it *at all.* Having a completely reliable understanding of all of the possible applications of a reason is not a necessary condition for having a perfectly satisfactory grasp of the reason's application, and a perfectly praiseworthy deliberative outlook, in any particular situation. This is for two reasons. First, the range of cases in which an agent understands a reason can be constrained without giving us reason to question his understanding within that range. Second, failures of sensitivity can often be explained by appeal to impairments of sensitivity in ways that are consistent with the correct exercise of the sensitivity in other cases where such impairments are not present. I will discuss such impairments in Section VI, below.

V. Degrees of sensitivity

Sensitivity to moral considerations often comes in degrees: agents can be perfectly sensitive to core cases and examples while being less sensitive, or even insensitive, to

more penumbral cases. I might, for example, be quite sensitive to the more obvious instances of injustice, but lack an appreciation of some of the more subtle and insidious ways in which injustice is perpetrated. So I may well be outraged that, on average, women are paid less than men for the same work, but see no problem with the fact that careers that are traditionally gendered feminine are paid less than careers traditionally gendered masculine: perhaps I believe that theses careers are now open to everyone, or that their respective rates of pay properly track merit. My outrage can issue from a genuine recognition of the injustice of gender inequality, and this recognition should not be called into question simply because I am unable to recognize more subtle version of the same injustice. My failure to extend my outrage to the full range of cases of injustice might well mean that I am not fully in possession of the virtue of justice, but an inability to recognize injustice in all its forms is not equivalent to an inability to recognize it at all. In fact, I may well be fully committed to the cause of equality, and I may work tirelessly for the cause of equal pay. It would be odd, and overly moralistic, to insist that, since I fail to see more subtle cases, I should not be praised for my commitment to injustice where I do see it. I can be faulted for my failure to extend my understanding of injustice without being faulted for my commitment to rectifying injustice within the range that I do, reliably, recognize.

Some periodic failures of moral sensitivity, then, reveal to us that the agent's understanding of morally relevant reasons is constrained to a more or less narrow range of cases, without thereby revealing any lack of sensitivity within that range. I may be perfectly reliably sensitive to the demand of justice within a narrow range, while being blind to its demands within a broader range. In practice, of course, this will reveal itself

in disagreements about what justice in fact entails: if I am truly committed to justice but my sensitivity to it is limited, *my* view will be that those who claim that justice requires an extension of pay equity are *mistaken* about what justice involves. Historically, moral progress has often involved expanding the scope of moral concern by bringing more and more cases out of the disputed or overlooked margins and into the clear central core.[21] So settling the question of whether my sensitivity is fully virtuous can require settling first-order debates about what justice in fact requires. Nevertheless, assuming that I am mistaken, and that justice does in fact require the extension of pay equity, my commitment to justice and equality in the narrow case may be quite sincere, and when I act on this commitment, my deliberative outlook may be perfectly appropriate. I may properly appreciate all of the relevant exclusionary reasons, recognizing, for example, that the fact that unequal pay would benefit me is of no consequence at all. Where the virtuous agent and I differ is in our deliberative dispositions regarding a broader range of cases: because I am not sufficiently sensitive the history of gender power imbalances in the workplace, I may fail to see that some facts about the wages historically commanded by different occupations are similarly irrelevant, since these differences in wages are the product of sexism and so reinforce current injustices. The virtuous agent and I can be in full agreement about the nature of the injustice in the narrow case, and the reasons to rectify it: our differences lie elsewhere, outside of the scope of this case. In other words, our difference concerns, not my deliberative outlook in the narrow case, but my deliberative disposition to extend my concern for justice to different and subtler cases.

[21] This point is frequently emphasized by Peter Singer, most notably in *The Expanding Circle: Ethics and Sociobiology* (Oxford: Oxford University Press, 1981).

My failure to extend my commitment to the cause of justice therefore reveals that I do not have the virtue of justice: I lack the appropriately virtuous deliberative dispositions, since I am not reliably sensitive to the reasons of justice in all contexts. Such failures can often be explained by the presence of false beliefs that prevent the non-virtuous person from forming correct judgments. These beliefs can be moral beliefs, but they can also be straightforwardly empirical: perhaps I have false beliefs about the gender distributions of various professions, or about the qualifications required for employment in several traditionally feminine professions. Not only can such false beliefs have important moral consequences by leading me to fail to recognize cases of injustice; they can also be the product of morally significant failures of practical reason, such as hypocrisy and self-deception, failures which are not simply *causes* of a lack of virtue, but *examples* of it.[22] So an explanation of someone's failure to appreciate the true force and role of a reason can shed light on other failures and imperfections.

It could be, for example, that I balk at recognizing that traditionally feminine jobs are underpaid because I enjoy having the relative status that comes with having a traditionally masculine job, and so my beliefs about the differences between traditionally feminine and masculine careers will be a product of self-interested self-deception. No doubt many such failures can be explained in this way. Nevertheless, it would be a mistake to suppose that, since my *failure* can be explained by appeal to the power of self-interest and self-deception, my apparent *success* should as well.

[22] I will discuss the moral status of hypocrisy and self-deception in Chapters 4 and 5, respectively.

It is true, of course, that some failures reveal to us that purported successes were only apparent: someone who never keeps promises that inconvenience him strongly suggests to us that, when he *does* keep a promise, it is because it he sees it as aligning with his own self-interest, rather than because he recognizes that his promise put him under an obligation to do so. Not all failures, however, force us to draw such conclusions. Many failures, such as my failure to fully extend my sensitivity to the reasons of justice, leave apparently successful exercises of that sensitivity untouched. Such failures might reveal that I am not as resistant to the forces of self-deception and the appeal of self-interest as I ought to be without showing that I am systematically self-deceived or motivated only by self-interest. Sensitivity to reasons, like resistance to failures of practical reason such as akrasia, hypocrisy, and self-deception, is not an all or nothing affair: it can come in varying degrees, and can be applied in wider and narrower ranges. Agents can be akratic, or hypocritical, or self-deceived, in some contexts but not in others: indeed, it only makes sense to talk of those who display such failures as genuine *agents* if they get things right at least some of the time. As a result, agents can adopt praiseworthy deliberative outlooks on their actions in some cases without reliably displaying such an outlook in all cases (and so without having fully virtuous deliberative dispositions), which is to say that the actions of non-virtuous agents can be morally worthy and deserving of praise, even if their actions are not reliably or consistently praiseworthy.

VI. Explanations for failure

Not all failures of moral sensitivity are evidence of a corrupt character and a complete failure of moral understanding, not only because an agent's sensitivity can be reliable and deliver praiseworthy deliberative outlooks within a narrow range without being equally reliable in wider contexts, but also because some failures of sensitivity are merely occasional lapses that admit of particular explanations that do not extend to other cases. This suggests that, in order to determine what an agent's failure of sensitivity reveals, we need to attend to the specific *explanation* of the relevant failure. When we do so, we will often see that such failures can be explained in ways that leave untouched the agent's claim to an appropriate sensitivity in other cases, either because the failure has a particular explanation, or because it occurs in a case that lies outside the range in which the sensitivity is appropriately reliable. What matters, then, is the nature of the explanation on offer.

McDowell frequent use of perceptual metaphors to explain the virtuous agent's sensitivity to reasons is helpful here.[23] After all, our *actual* perceptual capacities can be impaired in ways that do not call into question their sensitivity or reliability when the impairments are absent. Perception can be seriously impaired by fatigue, distraction, intoxication, too much sensory stimulus, and a whole host of other physical causes: our ability to distinguish the signal from the noise varies according to both our own physical and psychological state and the state of the environment in which we are trying to draw

[23] The self-controlled person "does not fully share the virtuous person's perception of the situation," while the virtuous person "has a reliable sensitivity to a certain sort of requirement," "sees" reasons that the non-virtuous person does not; and has the reason-giving force of some considerations "silenced."

such distinctions. Such impairments have an essentially cognitive aspect, as they involve, not simply an impaired ability to see or to hear, but an impaired ability to pay attention to, and properly form beliefs about, what we do see and hear.[24] But the fact that a migraine headache or severe intoxication can make it difficult to see well enough to read does not at all call into question my ability to see well enough to read when I am pain-free and sober.

Much the same point holds true for our sensitivity to *reasons*. After all, there is a sense in which McDowell's talk of perception is not metaphorical at all. Perceiving reasons standardly requires actual perception—sight, hearing, and so on—and, just as importantly, the cognitive capacity to properly process such perceptions, and these perceptual and cognitive capacities can be impaired by factors such as pain, fatigue, and intoxication. And, just as my inability to read with a headache does not in any way impugn the reliability of my ability to read while pain free, neither should my ability to recognize morally significant reasons for action be denied simply because that ability is similarly subject to impairment by factors such as pain, fatigue, and intoxication.

My ability to read is not called into question precisely because I am able to offer an explanation of my failure: I have a headache that is preventing me from focusing. The natural, and appropriate, reaction to my sudden inability to read is not "I suppose you couldn't really read all along", but rather "Your headache must be quite bad if it is keeping you from reading." Similarly, my general understanding of a morally relevant

[24] For a discussion of the effects of pain on attention, see e.g. Geert Crombez et al., 'The Disruptive Nature of Pain: An Experimental Investigation', *Behav. Res. Ther.*, 34 (1996), 911-18. For fatigue, see Maarten Boksem, Theo Meijman, and Monicque Lorist, 'Effects of Mental Fatigue on Attention: An ERP Study', *Cognitive Brain Research*, 25 (2005), 107-16.

reason should not be called into doubt simply because, whether because of fatigue, hunger, pain, or intoxication, I occasionally fail to recognize that the reason applies. Again, the natural reaction to my fatigue-induced failure to remember that I promised to do something should not be "I guess you don't understand the reason-giving force of promises", but rather "you must really be exhausted—you completely forgot that you promised." Apparently periodic failures of understanding only suggest that the understanding is lacking if there is no explanation on offer for the variation in performance. But, provided we can offer the right kind of explanation for *why* someone's understanding has been impaired, we need not conclude that his understanding is lacking in those cases where the impairment is not present. A sensitivity does not have to be 100% reliable for it to work when it does, just as the fact that my computer occasionally freezes gives me no reason to believe that it is not working as I type these words. What is important, for us to be confident in the deliverances of the sensitivity, is not that it be perfectly reliable, but only that we can provide an explanation of the sensitivity's failures in the reasonably small number of cases in which it is not reliable, and perhaps that we can provide a test for determining whether the conditions that prompt the failure obtain. And if we can, then non-virtuous agents—those who do not always do the right thing, and whose sensitivity to reasons is sometimes or even frequently impaired—can, on occasion, take the appropriate deliberative outlook on their actions, and so can act in morally worthy ways, so long as we can properly explain their occasional failures to do so.

Physiological factors such as pain, fatigue, hunger and intoxication are not the only candidate explanations for the cognitive and perceptual failures that lead to impaired

sensitivity to reasons. Such failures can also have more specifically psychological explanations. Empirical research on depression, for example, has shown that "mild to moderate depression… can be associated with quantifiable deficits in sustained attention of a similar order to those found in traumatic brain injury."[25] The same study showed that the attention of depressed subjects diminished after errors, suggesting that knowledge of the cognitive impairments brought about by depression exacerbated those impairments. Nor is depression the only cause of impaired sensitivity to reasons: distraction, stress, grief, anxiety, and even extreme joy can all make us less likely to notice reason-giving considerations than we otherwise would be. But the fact someone who is overcome by grief is not properly sensitive to the suffering of others—and so cannot see that suffering as providing him with any reasons for action—does not in any way entail that he is not sensitive to the suffering of others when he is not overcome by grief.

Once we admit that occasional failures brought about by specific and temporary impairments do not force us to doubt the reliability of an agent's sensitivity to reason in cases where such impairments are not present, there is less of a barrier to drawing a similar conclusion in cases where an agent's sensitivity is limited to a narrower range than the range demanded by a full possession of virtue. After all, if we allow for temporary impairments, we do not require that a sensitivity be reliable on every possible occasion in order for it to be reliable at all. Moreover, an agent's narrower than ideal range of sensitivity can itself be amenable to the same sorts of explanations that explain occasional impairments of sensitivity.

[25] Lydia Farrin et al., 'Effects of Depressed Mood on Objective and Subjective Measures of Attention', *Journal of Neuropsychiatry and Clinical Neuroscience,* 15 (2003), 98-104 at 103.

One important category of psychological explanation for impaired sensitivity to reasons is conditions such akrasia, self-deception, hypocrisy, complacency, self-interested rationalization, and a whole host of other failures of practical reason. These can explain why we failed to notice or be sensitive to reasons that, absent the impairment, we would certainly have noticed. My hypothetical failure to be properly sensitive to certain demands of justice, for example, can be explained by mistaken beliefs that are the product of self-deception, and my self-deception can, in turn, be explained by an excessive concern for social status or a self-interested regard for my own economic status.

The fact that someone's cognitive functioning is impaired when he is hungry gives us no reason for doubting that he is free of the impairment when he is well-nourished. Again, we can only make sense of his hunger *impairing* him he is not impaired when he is well fed. Similarly, the fact that my lack of sensitivity can be explained by self-deception should not make us doubt the possibility of my forming beliefs in an unbiased way in some other situation where the motives that led to my self-deception are not present, or are not as strong. It is easy for us to see that, in non-moral contexts, those who are self-deceived are not *always* self-deceived. Anyone who falls victim to self-deception must have a large store of true, unbiased, non-self-deceptive beliefs in order for self-deception to get off the ground, because he must have a large store of such beliefs in order for him to even count as an epistemic agent, capable of forming beliefs. The situation is no different when the beliefs in question are moral, and concern the force of certain considerations as reasons for action.

VII. Internal and external sources of impairment

A defender of the standard view might object that this last category of explanation—psychological explanations such as self-deception, akrasia, hypocrisy, moral complacency, and the rest—is of a different kind from the first, physiological category of explanation. The first category seems to be, in some sense, *external* to practical reason. Fatigue and pain do not have their home in practical reason, though they can impair it by interfering with our perceptual and cognitive capacities. Self-deception, hypocrisy, weakness of will, on the other hand, seem to be *internal* to practical reason: they are sources of cognitive impairment that emerge from within practical reason itself.

Perhaps this difference is significant. But how? One possible suggestion is that, while virtuous agents are no more immune than the rest of us to the kinds external sources of impaired practical reasoning, they *are* immune to the internal sources of such impairments. If so, then the fact that virtuous agents can be distracted while tired offers no support for the claim that non-virtuous agents can, on occasion, match the deliberative outlook of virtuous agents.

Before we evaluate the plausibility of this claim, we should note that it represents a significant retreat from the counter-factual condition of the standard view of virtue, since it allows that even the most virtuous agents will not be perfectly reliable. After all, even the most virtuous among us experience pain, hunger, and fatigue. When they do, their sensitivity to reasons is impaired: there is nothing about possessing a virtuous character that makes a person immune to the basic facts of human physiology. This presents the standard view of virtue with something of a dilemma. Either it has a threshold for virtue that ignores such facts, and is therefore so demanding that no one

could ever hope to meet it, or else it allows that virtuous agents are not *infallibly* reliable, since their sensitivity to reasons can, on occasion, be impaired by external factors. If the moral psychology of virtue is to be at all descriptively adequate to people as we actually find them, it has to take the second horn of the dilemma, and allow that the reliability required for virtue is not total.

If we choose this horn, and allow that even virtuous agents can fall prey to pain, fatigue, and hunger, the claim that virtue issues in a reliable sensitivity to reasons has to be modified to the claim that virtue issues in a reliable sensitivity to reasons, except for when that sensitivity is impaired by causes external to practical reason. This is a significant modification of the standard view of virtue. According to the current proposal, the virtuous person is no longer someone who is *always* and *reliably* sensitive to morally relevant reasons, and so who always has the right deliberative outlook on her actions. Rather, the virtuous person is someone who reliably has the appropriate deliberative outlook, unless her sensitivity to those reasons is impaired by an external factor that explains and excuses her failure. Even if this modification is correct, it involves retracting the claim that the virtuous person is someone who *always*, case-by-case, does the right thing.

In fact, even this modified version of the standard account will not do. It might seem plausible because it appears to reflect a distinction in blameworthiness: while agents should be blamed for failures of practical reason that have an internal source, they should *not* be held responsible for failures whose source is external. The thought here is that, say, akraisa is my fault, but being in pain is not. This strategy, however, will not work, since in fact I *can* be blamed for being tired, in pain, or hungry. We know full well

many of the physiological and environmental causes of the lack of attention and focus that can lead to impaired sensitivity to reasons, and we can very often avoid these causes if we choose. We can even *intend* to become tired, in pain, or hungry, or to be in environments that impair our cognitive capacities and our sensitivity to reasons. So we can certainly be responsible for our failures of sensitivity that arise from such physiological causes. The distinction gets even more problematic when we note that it puts psychological causes such as depression and grief on the internal, and so potentially blameworthy, side of the divide. These are explanations of impairment that are often taken to mitigate blame or even excuse us from it altogether, in part because they are impairments that almost no-one intentionally chooses. But self-deception and akrasia are similarly impairments that are rarely intentionally adopted: in fact, the forces of self-deception and the stresses on the strength of our will are often so strong that they can impair our sensitivity to reasons, and our judgments, even when we make an active effort to resist them.[26]

There is a second reason that this modified strategy will not work. The strategy relies upon a clean distinction between external and internal causes of impaired practical reason, but in fact these cannot generally be disentangled. Physiological factors such as hunger and pain are unlike purely physical causes of a lack of reliably virtuous behaviour, such as paralysis or injury. Someone who is paralyzed might recognize the reasons he would have to act in certain ways as readily as someone who is not: what is impaired is his capacity to *act* on those reasons. But factors such as fatigue and pain are

[26] I will have more to say about the moral significance of the non-intentional nature of self-deception in Chapter 4 and in particular in Chapter 5.

partly constituted by their cognitive effects, and so by their effects on our deliberative outlooks. They work in large part by paving the way for internal mechanisms such as akrasia, hypocrisy, complacency, and the like. To give but two examples: first, drunk people have a clearly external explanation for their cognitive impairment, and alcohol wreaks havoc on the brains of the virtuous and the vicious, the wise and the foolish alike.[27] But of course one of the defining characteristics of drunkenness is a kind of self-deception: drunk people believe they are funnier, cooler, smarter, more attractive, and, of course, *less drunk* than they really are, and it is seems clear that they often believe these things because they want them to be true. Alcohol creates the conditions for the forces of self-deception to take hold. Second, a very common form of akrasia is the failure to persist in one's firm intentions when confronted with pain, a phenomenon no doubt familiar to those who have abandoned a taxing exercise regimen. In impairing our ability to focus and process information, to stand firm in our judgments and to successfully carry out our intentions, factors such as hunger, fatigue, and pain make us more susceptible to failures of practical reason like akrasia, self-deception and hypocrisy. This means that an important aspect of virtue is an ability, not only to recognize and avoid situations that are likely to lead to pain and fatigue, but to *manage* pain and fatigue in situations in which they arise. This need not demand super-human strength and the ability to reason perfectly and keep one's calm while being tortured. It does, however, require that we are able, to some extent, to 'work through the pain': someone who becomes cruel and insensitive when struck with the most mild of headaches does reveal a certain lack of virtue, since

[27] Socrates is said to have been able to drink all night without showing signs of drunkenness, but in general the ability to resist the effects of alcohol tracks sex and body mass, not wisdom.

being able to deal with some degree of pain, fatigue, hunger, and the rest is part of having a stable and robust disposition to respond appropriately to moral reasons.

The distinction between internal and external sources of failure therefore cannot rescue even a modified version of the counter-factual condition. The conclusion to draw is that agents whose sensitivity to moral reasons is not fully reliable can nevertheless adopt the same morally worthy deliberative outlooks as those whose sensitivity is more perfectly reliable, provided that we can provide the proper explanation for their failures of sensitivity. Non-virtuous agents may have a diminished capacity to resist impairments of practical reason, and so they may not always have the right deliberative outlook and they may even fail to do the right thing. These facts alone, however, do not entail they can never do the right thing, or that they can never do so with a praiseworthy deliberative outlook.

VIII. Differences between virtuous and non-virtuous agents

If the preceding argument is correct, then flawed, imperfect, non-virtuous agents can, on occasion, act in ways that are just as praiseworthy as reliably virtuous agents. But this does not mean that there are no differences between virtuous and non-virtuous agents other than a statistical one in the probability of morally worthy action. In this final section, I consider some of those differences.

First, there are some worthy actions that are inaccessible to agents who lack the reliability of virtue: friendship, loyalty, constancy, and integrity are not virtues that can be approximated on rare occasions. It is constitutive of the reasons of friendship, for example, that they only occur within the context of an ongoing reciprocal relationship,

and so I simply cannot be a genuine friend if the fancy only strikes me once a year. Whatever reasons I act on when the fancy strikes me, they cannot be reasons of *friendship.* But this is a fact about the nature of the reasons of friendship (along with virtues such as loyalty, constancy, and integrity), rather than a fact about understanding reasons in general. The reasons relevant to courage, kindness, and justice, for example, are not constituted in the same way, and so are accessible much more infrequently. So while there are some moral reasons to which only virtuous agents can be sensitive, this in no way extends to all or even most moral reasons.

A second difference concerns the self-knowledge of virtuous agents as opposed to non-virtuous ones. The perfectly virtuous person is aware of her physical and psychological limitations, and so she makes efforts to avoid them and to minimize their effects: she gets plenty of rest, eliminates distractions, and avoids intoxication, and when she is tired, or in pain, or distracted, she reminds himself that this will have an effect on her sensitivity to reasons, and so perhaps she withholds judgement. The virtuous person knows her own limitations and takes them into account in deciding what to do and in carrying out her intentions. The possibility of such self-regulation partly explains why we can be responsible for failures of sensitivity brought on by external impairments. Such self-regulation requires a degree of self-knowledge, and failures of self-regulation often involve failures of self-knowledge.

Such self-knowledge is not confined to resisting physiological or external impairments: the virtuous person is equally aware of her susceptibility to failures of practical reason such as akrasia, hypocrisy, and self-deception. In the chapters that follow, I will explore in greater detail the ways in which such impairments involve

failures of self-knowledge, and argue that an honest and accurate self-assessment is an essential element of a stable and reliable sensitivity to moral reasons.

A third difference between virtuous and non-virtuous agents comes about as a result of their varying degrees of self-knowledge: because they have mistaken self-assessments, non-virtuous agents' sensitivity to reasons is fragile, or contingent, in some important ways. An agent's sensitivity is fragile if it only works in ideal conditions or if it can be easily impaired, whether by seemingly external factors such as fatigue and pain, or by more internal factors such as weakness of will and self-deception. An agent whose sensitivity to moral reasons is fragile in this way, and so is liable to break under psychic pressure, has a sensitivity that is contingent on things going well—on an absence of various psychic pressures, and so on the presence of more or less ideal conditions for understanding.

On the standard view of virtue, an agent who fails to be properly sensitive to moral reasons in some cases reveals that he completely lacks the proper sensitivity to such reasons in all cases. We should not, however, confuse the contingency of an agent's sensitivity with a *lack* of sensitivity in cases where he appears to get it right. It is tempting to suppose that a person who does not act virtuously under stress reveals to us that, like Kant's prudent grocer, he only acts in morally desirable ways when this fits with his own self-interest. But this is a mistake: non-virtuous agents can act for reasons other than self-interest, and can act in morally worthy and self-sacrificing ways. What makes them non-virtuous is not that they are only motivated by considerations of self-interest, but rather that their concern for morality is not as stable and consistent as it could be. Their lack of stability, in turn, is frequently a result of failures of self-

knowledge: failures to properly assess their own moral status and character, or to understand the motivations and implications of their own beliefs. They can increase the stability of their commitment only by increasing such self-understanding. Such agents do not withstand the various challenges to moral agency that life presents to all of us as well as they ought to—following the standard view's identification of virtue with practical wisdom, we might say that to be practically wise is for one's practical reason to be sufficiently robust to withstand the sorts of psychic pressures that cause the practical rationality of other less virtuous agents to fail.

The point to emphasize, however, is that to discover that a person's sensitivity to reasons is fragile and that the proper exercise of practical reasons is contingent is to discover a fact about the person's *character.* It is to learn that he is not fully reliable, which is to say, that he lacks self-knowledge, that he is prone to self-deception, that he cannot handle pain, or that he is weak in the face of certain sorts of temptation. But such facts about his character are general claims, and are not facts about the exercise of his sensitivity in each particular case. That someone's sensitivity is contingent is perfectly consistent with that sensitivity being properly sensitive in a wide range of particular cases. Such a sensitivity may not be *fully* reliable, since it is contingent on certain conditions obtaining. Nevertheless, we can be quite confident in the deliverances of the sensitivity, and we can properly call the sensitivity reliable, so long as we know those conditions obtain.

One reason that the sensitivity to reasons of so many people is fragile, contingent, and constrained is that stability, like sensitivity to reasons and self-knowledge, comes in varying degrees and is, inevitably, the result of a developmental process. Practical

rationality does not come fully formed: all of us are more or less rational, more or less of the time, in part because practical rationality is an ability, the development of which that takes time, effort, and training, and there is no guarantee that any of us will either attain fully stable practical rationality, or maintain it once we have attained it. If we are to make sense of moral education as a developmental process, in which we become increasingly sensitive to the force of moral reasons, and increasingly able to place them in the proper deliberative framework, then we should acknowledge that this process can be undertaken at different speeds and that it can have different end-points. Some attain the stability of virtue, others come close but fall sort, and still others attain it only to fall away again.

This suggests that we should abandon another of the central elements of the standard view of virtue: the claim that virtue is a state of character that one either possesses or lacks, and that we can therefore accurately categorize people as either virtuous or non-virtuous. None of us are perfectly rational deliberators, none of us are immune from false beliefs—including, crucially, false beliefs about ourselves—none of us can wholly escape fatigue, pain, and hunger, nor can we completely avoid grief, worry, or fear. We can be do better or worse at handling the effects these conditions have on our sensitivity to reasons, but it is simply unrealistic to expect that any of us is wholly immune to them. No matter how reliable we are, that reliability is contingent on avoiding a whole host of impairments over which we often have little or no control. In that sense, then, we are all imperfect agents: if virtue requires that, we at all times be fully sensitive to all of the reasons we have to manifest all of the virtues, and that this sensitivity not be contingent on the absence of any psychic pressures, then none of us are virtuous, since this is an impossible demand. But we should reject this account of virtue: some of us

have a greater sensitivity to morally significant reasons than others, and those who exercise that sensitivity reliably, if imperfectly, should be acknowledged as virtuous. As a result, we should reject the standard view's wide gap between 'virtuous' and 'non-virtuous' agents: virtuous agents exist on a continuum with the rest of us, and that the deliberative outlook of virtue is one to which even the most imperfect among us can aspire.

In this chapter, I have argued that it is possible, and even common, for the actions of morally imperfect people to be perfectly praiseworthy on some occasions even if they are deeply flawed on others. But it is consistent with what I have argued thus far that imperfect people can only merit praise on occasions in which no trace of their imperfections appear; that is, so far I have only considered cases where the relationship between praise and imperfection is diachronic. In the chapter that follow, I will argue that this relationship can also be synchronic; in other words, I will argue that actions that display morally significant failures of practical reason, such as akrasia, hypocrisy, and self-deception, can at the same time be morally worthy.

Chapter 3: In Praise of Akrasia?

Praise and blame are potential evaluations of an agent's intentional actions, and there is clearly a close conceptual connection between them. To be open to one is, in principle, to be open to both: there seems to be something conceptually confused about an agent who is, *in principle*, open to praise but not blame, or blame but not praise. But this does not mean that an particular *action* can be, in principle, both praised and blamed. Failures of practical reason such as akrasia, hypocrisy, and self-deception seem to be barred from praise: as *failures*, they can merit blame, but not praise. Not all instances of akrasia have moral import, of course, so not all cases are morally blameworthy. This does not mean that such cases cannot come in for *other* kinds of criticism, of course, only that they do not come in for moral criticism.[1] But there does seem to be an important asymmetry between praise and blame in the case of akrasia: since it is a failure, it can merit moral blame, but not moral praise.

In this chapter, I consider the plausibility of this asymmetry between praise and blame. I begin by setting out some reasons for thinking that akrasia can never merit praise. I then consider a case of 'inverse akrasia', that of Huckleberry Finn, and argue that Huck's actions are, at the same time, both akratic and praiseworthy. Finally, I consider some important differences between the deliberative outlooks of standard cases

[1] As Donald Davidson perceptively noted, akrasia "is not essentially a problem in moral philosophy, but a problem in the philosophy of action." Nevertheless, akrasia often *does* have moral import, even if we studiously avoid the unfortunate tendency to moralize about the pleasure of others. Donald Davidson, 'How Is Weakness of the Will Possible?' *Essays on Actions and Events* (Oxford: Clarendon Press, 2001) at Note 14.

of akrasia and praiseworthy cases of apparent akrasia. Regardless of whether cases like Huck are properly called 'akrasia', my argument is that in moral contexts, an agent's actions can be praiseworthy despite being characterized by a deliberative outlook that features significant failures of practical reason, including something very like akrasia.

I. Praise and blame: acting for a reason

Intentional actions are just those actions for which we can be held responsible. Since intentional actions are those that are done *for reasons*, there is a close connection between assessments of praise and blame and assessments of rationality. As I argued in Chapter 1, praise is appropriate only for intentional actions done for a particular *sort* of reason and with the right deliberative outlook on that reason. This thought leads Kant, for example, to deny that there is any moral worth in the actions of the honest shopkeeper who, out of self-interest, never cheats his customers. And it is reflected Aristotle's distinction between states "merely… in accord with the correct reason" and those "involving correct reason."[2] Merely doing the right action for *some* reason is not enough: praise can be reserved for those who do the right action for the *right reason*, and with the right deliberative outlook, just as blame is withheld from those, such as addicts, whose actions are done *for no reason* at all.

The same seems to hold true of blame: those who act for morally *inappropriate* reasons are subject to blame just as those who act for the right sort of reason are praised. Likewise, we can be blamed for various rational failures. To blame someone for negligence or carelessness, for example, in not to criticise her particular reasons for

[2] Aristotle claims that only the latter is virtue. Aristotle, *Nicomachean Ethics* at 1144b28.

action, but rather to criticize her failure to recognize the situation as one that required greater care or attention. This too is a rational criticism: a negligent person does not properly appreciate the reasons she has for taking more care, and is blameworthy because she *should have* appreciated them.

This line of argument reflects powerful intuitions about the connections between praise, blame, and rationality, and gives voice to the idea that it would make poor sense of the importance we attach to morality if it were possible to be irrational and yet merit praise, or to be fully rational and yet merit blame.[3] This is because morality is not a distinct domain from practical reason, as if there were a domain of moral considerations that were wholly distinct from rational considerations. Quite the contrary: morality and practical reason are inextricably bound up. Assessments of moral responsibility, in turn, are in part assessments of us as rational agents. That helps explain why we are neither praised or blamed for actions that are genuinely unintentional, compelled, or un-chosen. To say that someone merits praise is to say that she was moved to act by the appropriate reasons, and to say that she merits blame is to say that she failed to do so, either because she failed to recognize the correct reasons or because she acted for the wrong one (or both).

In the first two chapters, I emphasized the close connection between, on the one hand, an agent's reasons for action and his deliberative outlook, and, on the other hand, the praiseworthiness of his action. This might seem to involve an overly intellectualized and rationalistic view of praise, as if an agent only merits praise if he acts in full

[3] This second claim is more controversial. For a powerful argument that rationality cannot require us to act immorally, see Philippa Foot, 'Rationality and Virtue', *Moral Dilemmas* (Oxford: Clarendon Press, 2002), 159-74.

conscious awareness of the reasons that make his action right and without displaying any practical irrationality. Such a view would rule out praise for those agents who are not particularly self-aware, or whose practical rationality is not perfectly developed, and so would seem to needlessly rule out praise for a great many of us. In fact, though I do emphasize the importance of an agent's deliberative outlook, I reject overly intellectualized understandings of moral worth. In this chapter, I argue that it is possible for akratic actions to merit a considerable degree of praise. While an agent's reasons for action and deliberative outlook determine whether his actions should be praised or blamed, this need not rule out praising actions that display significant failures of practical rationality, including a lack of self-knowledge, and so the account of praise I defend is neither overly intellectualized nor exclusionary.

Acting for a reason does not require conscious deliberation prior to action. Nor does it require that the action be accompanied by an occurrent thought of the form "I do A for reason *x*."[4] Actions count as rational when the agent can justify them, by appealing to the reasons that make the action good. Acting for a reason requires the possibility of rational justification, rather then the presence of conscious deliberation.[5]

---

[4] An extended argument can be found in Nomy Arpaly, *Unprincipled Virtue* at 51-60. One interesting example she mentions is deliberation itself: while it can clearly be rational to deliberate about what we have most reason to do, we do not generally deliberate in order to decide that we should deliberate. If we did, an infinite regress would loom over the suggestion that deliberation is required for rationality and threaten the rationality of deliberation itself. More amusingly, Bernard Williams wonders what those who believe that all intentional action was deliberated action must seem like to their sexual partners. Bernard Williams, 'Voluntary Acts and Responsible Agents', *Making Sense of Humanity* (Cambridge: Cambridge University Press, 1995), 22-34 at 23.

[5] The classic account of this position is Elizabeth Anscombe's. She describes intentional acts are those to which "a certain sense of the question 'Why?' is given application. The sense… is that in which the answer, if positive, gives a reason for acting." G.E.M.

It is important, however, to distinguish general accounts of practical reason from accounts of praise and blame. Not all intentional actions involve prior deliberation or occurrent thoughts, but perhaps all *praiseworthy* actions do. On this view, actions like climbing the stairs and turning right on a red light might be intentional actions, but such non-deliberative intentional actions can have no positive moral worth. If the represent cases of negligence, they may merit blame, but they cannot be praised. Praise is reserved for moral actions carried out *because they are moral*, and so for actions that are accompanied by the thought that 'this action is morally appropriate, and I do it for that reason.'[6]

It is likely that no one actually believes that praise is reserved for actions that are the result of moral deliberation and accompanied by an occurrent thoughts about their rightness. Certainly few people would assent to the view when phrased in this way. It is nevertheless important to mention, since talk of deliberation is so prevalent in the philosophical literature, and particularly in moral psychology. Indeed, on the account of moral worth I have been developing, praise depends on an agent's *deliberative* outlook. But that does not mean that such an outlook necessarily involves conscious reflection. Conscious moral deliberation and occurrent beliefs about morality can certainly be

---

Anscombe, *Intention* (Cambridge, MA: Harvard University Press, 1957) § 5**.** See also Joseph Raz, *Engaging Reason* (Oxford: Oxford University Press, 1999), especially chapter 2. The account of rational action given in Donald Davidson, 'Actions, Reasons, and Causes', *Essays on Actions and Events* also denies that deliberation or occurrent thoughts about reasons are required, though it differs in important ways from the account given by Anscombe and others.

[6] This claim is neutral between the various ways of interpreting this thought that an action is morally appropriate or required: Kantians might interpret is as a thought about what duty requires, Aristotelians as a though about the 'fine,' or virtue, or some *particular* virtue, and so on.

important, and the way we reflect on our reasons for action is worthy of philosophically attention. We should not think, however, that such conscious reflection is necessary for moral assessment.

One clear reason for rejecting the view that praise requires conscious deliberation is that actions that are the result of such deliberation can be *for that very reason* less praiseworthy than more spontaneous actions.[7] This is a point repeatedly emphasized by those who argue that character and attitudes as well as actions are morally significant. A person's character is in part expressed in the considerations that strike her as obviously significant, the considerations that strike her as irrelevant, and the considerations that simply do not occur to her. As Bernard Williams puts it, "some concerns are best embodied… in deliberative silence," by never being entertained at all.[8]

The moral worth of an agent's actions can be assessed without those actions being the result of conscious deliberation or occurrent thoughts. It is still unclear, however, what sort of justification is required for an act to merit praise or blame: what is required for the question 'why?' to be answered by giving the relevant sort of reason? The narrowest possibility is that the agent must be able, if asked, to actually *answer* the question 'why?' by citing her reasons. She does not need to have been considering those

---

[7] See, for example, Bernard Williams' famous argument that we can be blamed for having "one thought too many." Bernard Williams, 'Persons, Character and Morality' at 19. Scanlon makes a similar point with, like Williams, a husband and wife example: a husband for whom his wife's feelings do not "present themselves to him spontaneously as reasons", so that he has to "carefully monitor himself—reminding himself to go back and consider how his wife would feel" exhibits a fault, one that a husband for whom his wife's feelings present themselves spontaneously as reasons does not. T.M. Scanlon, 'Reasons and Passions', in Sarah Buss and Lee Overton (eds.), *Contours of Agency: Essays on Themes from Harry Frankfurt* (Boston: MIT Press, 2002), 165-83 at 172.

[8] Bernard Williams, *Ethics and the Limits of Philosophy* at 185.

reasons at the precise moment of action, but she does need to be able to identify them after the fact. She can do this, perhaps, because she already believed, prior to acting, that the reasons were good ones, and so she does not need to reconsider them at the moment of action.

Consider an analogy with perception. We all believe that vision is generally reliable, so we do not consider, each time we perceive something new, whether we in fact have good reason to believe that it exists. The reliability of visions does not enter our thoughts at all. If asked to justify the belief that the cat is on the mat, however, most of us would appeal to our perception of a cat on a mat, along with our longstanding belief that perception is reliable.

If this is the right account of justifiability, it involves two conditions: (1) the agent must believe (in some sense) prior to acting that her reasons are good ones, and (2) she must be able to explain those reasons after acting. *Neither* of these conditions, however, is necessary for the agent to have acted for a reason or for the agent to merit praise or blame.

Being able to explain one's reasons does not seem necessary for some clearly intentional actions, such as fast action in sports. A hockey player who chooses to shoot rather than pass on a 2-on-1 need not deliberate in order for us to either count his action as intentional or for us to assign it praise—'smart play!'—or blame—'you should have passed!' Part of being a skilled athlete is being able to make the right play without thinking. But why should our assessment of his action depend in the player's ability to explain his reasons after the fact? How articulate the player is does not seem relevant to whether his action merit praise or blame—what *is* relevant is the reason he made the play.

It may be that the player shot because he saw the goalie overplaying the pass, but that he is not able to identify this as his reason; all he might be able to say afterwards is 'I don't know, shooting just *felt* like the right play.'[9] This response does not *identify* his reasons; at most, it simply indicates that he was in fact moved by the reasons. Praise should not depend on how articulate the agent is, as anyone even casually familiar with professional athletes can attest.

While perhaps not articulate enough to fully identify his reasons for acting, the fast-acting hockey player might nevertheless have agreed, prior to the game, that it is best to shoot on a 2-on-1 when the goalie overplays the pass. In fact, it is this belief, instilled in him through years of practice, which allows him to recognize that he should shoot, even if the play is so fast that he cannot quite explain it afterwards. So perhaps the justifiability account requires that the agent believed, prior to acting, that the reasons for which he acted were good ones, and that he would have endorsed them as good reasons if asked prior to the moment at which he acts.

This suggestion, however, also excludes many legitimately praiseworthy intentional actions. It is possible for an agent to act for reasons that he only discovers he endorses when forced to consider them: some situations can reveal to us that our values are not what we thought they were. An excellent example is Lord Fawn, from Anthony Trollope's novel *The Eustace Diamonds*, discussed by Harry Frankfurt in "Rationality and the Unthinkable." Lord Fawn asks Andy Gowran, a lower-class estate steward, to give him scandalous gossip about Lord Fawn's fiancée. In the midst of the conversation,

[9] As Anscombe puts it, the question 'Why?' "is not refused application" when the answer is "no reason" or "I don't know why I did it." Anscombe, *Intention*. § 17.

however, Fawn discovers that he simply cannot continue: "Every feeling of his nature revolted against the task before him…He was weak, and foolish, and in many respects ignorant, but he was a gentleman… He paused for a moment, and then he declared that the conversation was at an end."[10] In Frankfurt's terms, Lord Fawn's feelings revolt against his planned course of action. Frankfurt's claim, however, is that in revolt, Lord Fawn's feelings are *not* "opposed to the agent. They are in the most authentic sense his own forces, which are… integral to his nature."[11] Lord Fawn discovers, when confronted with the reality of the situation, what he truly judges he ought to do.[12]

Lord Fawn abandons his intended course of action under the influence of his emotions. Nevertheless, Lord Fawn is not akratic. He has an extreme aversion to acting as he thought he could or should, something that is also true of many akratics. But in addition, says Frankfurt, "the aversion has his endorsement." This endorsement sets agents like Lord Fawn agents apart from those whose "inability is due to addiction or some other irresistible impulse." [13] It also, however, sets Lord Fawn apart from akratics: after acting, he endorses his reasons for action and sees his actions as justified, even though he would not have done so prior to acting. Though his feelings lead him to abandon the course of action he had decided upon, his actions are not akratic, since those feelings have his endorsement.

---

[10] Cited in Harry Frankfurt, 'Rationality and the Unthinkable', *The Importance of What We Care About*, 177-90 at 183.
[11] Ibid. p. 184.
[12] Bernard Williams, in discussing cases of moral incapacity such as Lord Fawn's, writes that an incapacity can "present itself at once to the agent as a decision, and as a discovery." Bernard Williams, 'Moral Incapacity', *Making Sense of Humanity*, 46-55 at 52.
[13] Frankfurt, 'Rationality and the Unthinkable'. p. 182.

The difference between Lord Fawn and the akratic agent has, in part, to do with regret. Regret is characteristic of akrasia, since akratic agents continue to believe, after acting, that they acted against their own best judgment. Cases like Lord Fawn's, however, need not involve regret at all. In endorsing his feelings, and so his decision to act on them, Lord Fawn need not experience any regret.[14] Quite the opposite: he sees his actions as appropriate, even if he did not see them as such prior to acting.

In endorsing his feelings, Lord Fawn acts for reasons. Frankfurt puts the point, somewhat misleadingly, as follows: "feelings may accord better with reason that judgment does."[15] Perhaps a better way of phrasing this is to say that feelings may accord with reason better than belief does, and can reveal one's genuine judgment. For Lord Fawn, in the end, judges that he cannot dishonour himself and his fiancée by continuing his conversation with Gowran, whatever he might have believed beforehand. This judgment was formed on the basis of his emotional reaction to the situation, an emotional reaction that was a "manifestation of [his] fundamental rationality."[16]

Characters such as Lord Fawn, who are led by their emotions to act against their prior intentions, can clearly be responsible for their actions, and can merit both praise and blame, particularly if they endorse their reasons after acting. The particular case of Lord Fawn is ambiguous, since it is unclear whether we should praise him for his gentlemanly

[14] He may regret the original decision to have the conversation, or even that he has a mistrustful character, but he will not regret his actions in ending the conversation. This is quite unlike the akratic agent, who *would* regret those actions.
[15] Frankfurt, 'Rationality and the Unthinkable' p. 189.
[16] Ibid. p. 190.

refusal to sully his fiancée's honour, or blame him for his class-prejudiced snobbery.[17] What is clear, however, is that acting on the basis of one's emotional reactions and contrary to one's previous intention does not exclude one from responsibility, and even leaves it open that one's actions merit praise.

It seems, then, that both praise and blame can attach to actions that are done without prior deliberation or occurrent thoughts, to actions that the agent has difficulty justifying after the fact, and to actions whose justifications the agent rejected prior to acting, and only endorsed at the time of action. In all of these cases, the agents act for a reason, but (in different ways in each case) they do not clearly identify the reason for which they act. In each case, however, the agents also in some sense endorse the reasons for which they act. Even the inarticulate agent is able to endorse his reasons for acting if they are put to him—his inability to identify them does not mean that they do not have his endorsement. So long as agents can offer or endorse their reasons for acting, then, their actions can be subject to moral assessment.

If that were correct, then it would explain why there is an asymmetry, in the case of akrasia, between praise and blame. In all of the many possible classes of action where both praise and blame can potentially be assigned, the agent in some sense endorses or accepts her reasons for acting. This endorsement may be a longstanding product of conscious deliberation, or it may be discovered through an emotional reaction. But since

[17] Frankfurt mentions a more clear-cut case of praiseworthiness: military officers who refused to carry out orders to launch a nuclear weapon (they did not know it was a drill). Since the officers volunteered for their assignments, their refusal was not based on a previous decision—they had no doubt affirmed many times their willingness to launch nuclear weapons. Rather, "they discovered that participating in the initiation of a nuclear assault was for them unthinkable." Ibid. p. 182.

the endorsement is present, the actions can be assigned both praise and blame. Akratic agents, however, do *not* seem to endorse their reasons for acting, and this might rule out the possibility that they can be praised.

II. Praising akrasia?

Neither Lord Fawn nor the inarticulate hockey player is necessarily irrational, even though they do not have full deliberative access to their reasons. Akrasia, however, is different. It involves a level of irrationality that the other cases do not. Aristotle describes the akratic as someone who "because of his feelings abandons himself against correct reason."[18] Davidson perceptively points out that akrasia need not involve *desire* overcoming judgment, but he does keep the importance of acting against correct reason: his discussion opens with the claim that "an agent's will is weak if he acts, and acts intentionally, against his best judgment."[19] Akratic action is therefore a paradigm case of irrationality, since akratic agents are irrational by their own lights: they do not see their own actions as rational. [20]

---

[18] , 1151a21.

[19] Davidson, 'How Is Weakness of the Will Possible?' p. 21. A similar definition can be found in Alfred Mele an action is akratic if it is "*free, intentional* action contrary to the agent's better judgment" or, more accurately, his "decisive better judgment." Alfred Mele, *Irrationality* (Oxford: Oxford University Press, 1987) at 4-6. See also David Wiggins, 'Weakness of Will, Commensurability, and Objects of Desire', in A.O. Rorty (ed.), *Essays on Aristotle's Ethics* (Berkeley: University of California Press, 1980) at 241. "Anyone not under the influence of a theory will say that, when a person is weak-willed, he intentionally chooses that which he knows or believes to be the worse course of action when he could choose the better course."

[20] This is the standard view of akrasia. It is not universally shared, however. Perhaps the earliest philosophical discussion of akrasia is in Plato's *Protagoras*, where Socrates argues, in effect, that akrasia is impossible, since no one ever knowingly chooses to do wrong. All apparent cases of akrasia are in fact cases of ignorance. Plato, *Protagoras*,

Despite this irrationality, however, akratic actions are free and intentional actions. In other words, they may be irrational, but they are done for reasons. Akratic agents do not act for no reason at all: on the standard interpretation, they act for a reason, just one that they do not judge to be good or sufficient for action. Because they act freely and intentionally, it is standardly believed that akratic agents can be held responsible. This is one thing which separates akrasia from compulsions such as addiction: akratic agents could have acted otherwise, while compelled agents could not, since the latter, but not the former, acted on the basis of an arational and irresistible urge. So responsibility seems built into the very possibility of akrasia: to say that someone acted akratically, as opposed to compulsively, is to say that the former, but not the latter, could have exercised self-control and so is responsible for her actions. This assessment of responsibility seems to run in one direction only: in moral contexts, akrasia can be blameworthy, but it cannot be praiseworthy.

But is this right? Can an apparently akratic agent never merit praise, regardless of how they act? Akratic agents are standardly represented, in moral contexts at least, as knowing how they ought to act, and yet being led to act immorally by their desires. Thus Aristotle: akrasia "is similar to vice in its actions"[21] and the akratic is "someone who

---

trans. C.C.W. Taylor (Oxford: Oxford University Press, 1996). For a different sceptical argument, which does not dispute the existence of akrasia but does question the distinction between akrasia and compulsion, see Gary Watson, 'Skepticism About Weakness of Will', *Agency and Answerability* (Oxford: Clarendon Press, 2004), 33-58.Watson does not reject the distinction altogether, but he does argue that it is more "relativistic" than the standard view allows. Another dissenting view can be found in Richard Holton, 'Intention and Weakness of the Will', *The Journal of Philosophy*, 95 (1999), 241-62.

[21] Aristotle, *Nicomachean Ethics*. 1151a6.

because of his feelings abandons himself against correct reason."[22] But recall Davidson's point that akrasia is fundamentally a failure of practical reason, and is not always a specifically moral failure. It would seem at least possible for someone to act for a reason that he did not endorse, but for this akrasia to be 'similar to virtue in its actions.' Such an agent would 'because of his feelings abandon himself against incorrect reason.' The question, then, is twofold. First, is such "inverse akrasia" possible?[23] Second, if it is possible, can it be praiseworthy?

Inverse akrasia requires, at a minimum, doing a morally right action (not simply one that is not wrong), and doing so while displaying the same pattern of practical reasoning as standard akrasia. It may be that this is impossible, as it would be if no one who displays the rational failure of akrasia could even do a morally right action. But unless morally right actions are simply *defined* in such a way so as to rule out akrasia, this seems implausible. Akratic agents act for reasons, even if they are irrational in so doing. So even if it is insisted that morally right actions are those done for a certain sort of reason, it remains possible that inverse akratic agents act for those reasons, albeit while reasoning akratically, and so perhaps without a praiseworthy deliberative outlook.

Another suggestion is that what appears to be inverse akrasia does not represent a failure of rationality, since it would involve doing the morally right action, and the

[22] Ibid. 1151a21.
[23] The term "inverse akrasia" was first used to describe such cases in Nomy Arpaly, 'On Acting Rationally against One's Best Judgment', *Ethics*, 110 (2000), 488-513.

morally right action is the rational action. Since there is no failure of rationality, there is no akrasia. This appears to be Aristotle's strategy.[24]

There is, as we shall see, some truth to this second claim, but as it stands, it will not do. First of all, as Davidson rightly pointed out, the problem of weakness of will is not necessarily a moral problem at all, and so the rational failure can in principle persist regardless of the moral stakes. Secondly, the source of the problem is not that the akratic agent acts in a way that is, in fact, irrational: the problem is that he acts in a way that *he himself* deems irrational, and this conflict between what he judges he should do and what he in fact does can be present regardless of whether his actions are ones that he in fact has most reason to do. It may be that, since he has a good reason for doing what he did, the inverse akratic agent is not irrational in the same way that the standard akratic agent is, but inverse akrasia cannot simply be explained *away* by claiming that there is no failure of rationality at all.

Inverse akrasia, if it exists, will display many of the features of standard akrasia. There may be important differences, but it will still involve acting for reasons that the agent does not endorse or believe to be the best available. If it can be praiseworthy, it will be so *despite* displaying an important rational defect.

This question is best considered by examining an actual case, to see if inverse akrasia can be both coherently described and assigned moral merit. If it can, then the

[24] Aristotle mentions Neoptolemus, from Sophocles' *Philoctetes*, who cannot bring himself to tell a lie, despite being order to do so by Odysseus. Though Philoctetes' pain at lying prevents him from acting as he believed he ought to, Aristotle denies that he is akratic: "not everyone who does something because of pleasure is… incontinent, but only someone who does it because of a shameful pleasure." Aristotle, *Nicomachean Ethics*. 1151b24. Aristotle's denial obscures the similarity between inverse and standard akratics, since both act contrary to their own judgment.

asymmetry between praise and blame discussed above is only apparent. If not, then the asymmetry does indeed exist.

Huckleberry Finn

The most common example of inverse akrasia in the literature is Mark Twain's Huckleberry Finn. Huck is a young, uneducated boy from antebellum rural Missouri, and he has many of the values and beliefs common to that time and place. In particular, he does not question the moral justifiability of slavery, and he believes that slaves can and should be treated as the property of their owners. During the course of the story, however, Huck befriends Jim, a slave, and helps him escape. This action goes against Huck's strong belief that he ought to turn Jim in, since, as a slave, Jim is someone's lawful property. On two separate occasions, however, Huck is faced with an opportunity to turn Jim in, and on both occasions, he finds that he cannot, despite his belief that it would be the right thing to do. This causes him to feel intense regret; he berates himself for aiding in what he considers to be theft, and believes that he has a duty to return Jim to Miss Watson, Jim's owner. After all, what has Miss Watson ever done to Huck to merit stealing from her? Far from believing that he acted rightly, his conviction that he has repeatedly acted both weakly and badly convinces him that he is destined to remain a bad boy.

If Twain's description of Huck's psychology is coherent, then it seems as if something very much like inverse akrasia is possible. Huck displays a failure of practical reasoning similar to that of the standard akratic agent: he acts contrary to his beliefs about how he should act, and if he does act for a reason, it is not one he appears to endorse.

Moreover, his actions are at least in conformity with the demands of morality—I take it as uncontroversial that freeing slaves is, morally speaking, a very good thing. Twain clearly intends for the reader to think of Huck's actions as praiseworthy, and to see Huck as, essentially, a boy with a good heart, even if he has some deplorable (but, given his environment and upbringing, understandable) beliefs. Huck is Twain's creation, of course, but literary theory teaches us that we are not bound by to the author's intention when it comes to interpreting his works. Though Huck has acted in a morally desirable way, he might not merit any praise for his actions.

Jonathan Bennett, who first raised the case of Huck in this context, argues that Huck acts on the basis of what he calls "sympathy," which is distinct from moral judgment. Sympathy can sometimes be in accord with morality, and it can sometimes conflict, but, on Bennett's view, a person's sympathy is one thing, and his or her morality is quite something else.[25] Huck firmly believes that, in helping Jim escape, he has acted immorally—as Bennett perceptively points out, Huck does not even seem to weight the reasons for and against freeing Jim. His belief that freeing Jim is wrong is so firm that Huck does not seem to consider *any* reasons for rejecting this belief. *All* of the reasons that Huck consciously considers line up on one side: in favour of turning Jim in. In fact, he takes his inability to turn Jim in as evidence that he is a very bad boy, beyond redeeming.

[25] Bennett claims "*feelings* must not be confused with *moral judgments*." The former "is not a judgment about what I ought to do but just a *feeling*." Jonathan Bennett, 'The Conscience of Huckleberry Finn', *Philosophy,* 49 (1974), 123-34 at 124. Emphasis in original.

Far from giving reasons to justify his actions, Huck says several times that he does not really know what he is doing. “It hadn’t come home to me, before, what this thing was I was doing…” “Thinks I, this is what comes of not thinking.” At the crucial moments when Huck acts in ways that help Jim, he describes himself as acting weakly, rather than on the basis of any kind of decision, even an ill-considered one. Huck sees himself as lacking resolve, or “tuck.” Faced with a chance to turn Jim in, he laments that “he warn’t man enough” to do the right thing, that he “hadn’t the spunk of a rabbit” and could feel himself “weakening.” In his weakness, he tells a lie that helps Jim escape.[26]

Since Huck clearly believes that he ought to turn Jim in, never seems to consciously consider any reasons for helping Jim to escape, and experiences his actions as the result of weakness and lack of resolve, Bennett concludes that Huck acts merely on the basis of an “unreasoned emotional pull.”[27] If assessments of praise and blame are only appropriate for actions that are done for a reason, then it seems that, since he does not appear to act for a reasons, praise cannot be appropriate in cases like Huck’s. Inverse akratics, like standard akratics, might deserve blame for their failure to properly exercise their rationality (or their failure to develop their capacity to do so), but they cannot merit praise.[28]

---

[26] Mark Twain, *The Adventures of Huckleberry Finn* (New York: William Morrow, 1994). Chapter 16.

[27] Bennett, 'The Conscience of Huckleberry Finn'. p. 127.

[28] Bennett makes his view clear by comparing Huck to Heinrich Himmler, who apparently suffered the occasional pangs of sympathy for the victims of his death camps. The difference between Huck and Himmler, says Bennett, is simply that, in the struggle between unreasoned sympathy and bad morality, the victory went to sympathy in Huck’s case, and bad morality in Himmler’s. That Huck lacked the courage of his convictions is fortunate, but it is hardly something for which he should be praised. With a bit more strength of will, Bennett suggests, Huck would be just like Himmler. Ibid. p. 128.

Bennett clearly lays out the problem with seeing Huck's actions as praiseworthy. His actions do not seem to reflect his beliefs about how he ought to have acted, and this leads to the thought that his doing the right thing was quite accidental. We would not praise Miss Watson, Jim's owner, for allowing him to escape by simply forgetting to lock the door to their cabin, since she would have in no way intended to set him free. But if Huck thought that he had a duty to turn Jim in, and if he feels regret at failing to do so (just as Miss Watson might chide herself for failing to lock the door), why should we suppose that his actions are worth praising? There is a distinction, familiar from debates about consequentialism, between approving of the consequences of someone's actions and approving of the actions themselves. Bennett's argument is that, if we praise Huck, we miss the importance of this distinction. We might be glad that Jim has been freed, but this should not commit us to praising Huck any more than it would commit us to praising a forgetful Miss Watson.

Bennett's argument cannot simply be that Huck's freeing Jim did not come about as a result of conscious deliberation, or that he was not able to justify his actions by citing his reasons after the fact. As we have seen, meeting these conditions is not a necessary condition for an action being either intentional or praiseworthy. Bennett's argument must be that, since Huck's actions were the merely the result of an unreasoned emotional pull, they cannot be justified *at all*. Huck is not merely inarticulate: if asked, he would actually *deny* that freeing slaves was justifiable. It would be different if his sympathy and his judgment coincided, since then he could appeal to the reasons grounding his judgment in order to justify his actions. But since Huck does not endorse the emotions that lead him to

act, and sees then as signs of weakness, his actions cannot be rationally justified, and hence cannot be praised.

The first challenge is to show that Huck's actions can be explained by pointing to a *reason*, and not simply by claiming that he is weak. He may *also* be weak, but if he merits praise, then there should be a reason, other than cowardice, for his actions. This is the approach taken by Nomy Arpaly.[29]Arpaly interprets Huck as having undergone an unconscious "perceptual shift." During their time together, Huck has come to see Jim as a person, rather than as a piece of property. This explains Huck's reluctance to turn Jim in, "even if his conscious mind has not yet come to awareness of this perceptual shift." In acting in unconscious recognition of Jim's personhood, Huck "*is* acting for morally significant reasons. This is so even if he does not *know* or *believe* that these are the right reasons."[30] Her argument, in short, is that Huck in fact believes that Jim does not deserve to be enslaved; he just does not know that he believes this, because his belief is unconscious.

Arpaly takes Huck's praiseworthiness to demonstrate the more general point that "*sometimes an agent is more rational for acting against her best judgment than she would be if she acted in accordance with her best judgment.*"[31] Huck is *more* rational in

[29] Arpaly, *Unprincipled Virtue*. Chapter 2. Julia Driver offers a different account: Huck is praiseworthy for helping Jim escape because he acts from virtue. For Driver, this means that he acts from a trait that reliably produces good and significant social benefits. So Huck's praiseworthiness is independent of his reasons for action or his rationality: on Driver's this view, it could be his weakness or his squeamishness that constitutes his virtue. Aristotelians would of course reject this account of virtue. Julia Driver, 'The Virtues and Human Nature', in Roger Crisp (ed.), *How Should One Live? Essays on the Virtues* (Oxford: Clarendon Press, 1996), 111-30.
[30] Arpaly, *Unprincipled Virtue* at 76-7.
[31] Ibid. at 36, emphasis in original.

freeing Jim than he would have been if he had turned Jim in. The danger with this interpretation is that, in classifying Huck as rational, the force of the claim that he is akratic is weakened. Huck's actions may be rational, but it would seem that he still demonstrates an important failure of rationality: he acts for reasons that he does not endorse, even after acting. There is a tension in the claim that inverse akratics can merit praise, since we want to be able to explain the praise without explaining *away* the akrasia.

Arpaly's point is that Huck's actions do meet a minimal standard of rationality, and that he would be *even less* rational if he had turned Jim in.[32] Both Robert Audi and Alison McIntyre make similar claims. Audi argues that akratic actions can be rational because rationality should be understood holistically: the rational action is the one that is favoured by the balance of reasons that the agent has. Since we are imperfect deliberators, there is difference between the balance of reasons we have and the judgements we make on the basis of our reasons.[33] Alison McIntyre makes a similar

[32] Harry Frankfurt makes a similar point: "Whatever Hume says, to regard the destruction of the whole world as less important than a scratched finger is not a rational option. *It is lunatic*." Frankfurt, 'Rationality and the Unthinkable', at 185. It is sometimes claimed, narrowly, that irrationality requires inconsistency between beliefs, or between belief and action. Speaking more broadly, however, we often say that there are some beliefs or preferences that are *themselves* irrational. For an account of the distinction between narrow and wide irrationality, as well as an argument in favour of the narrow account, see Scanlon, *What We Owe to Each Other* at 25-30. As we shall see, if Huck had turned Jim in he would *still* have faced charges of inconsistency between his actions and (some of) his beliefs.

[33] Robert Audi, 'Weakness of Will and Rational Action', *Australasian Journal of Philosophy*, 68: 1990), 270-81.

point: "what actually moves us may… be evidence about what we could be moved to understand and justify through further reflection."[34]

This line of argument draws attention to an important distinction between an agent's beliefs about what he has most reason to do and what he in fact has most reason to do. An agent can be wrong about what he has most reason to do even if those reasons are understood subjectively or internally. Huck does not simply have more reason to free Jim than to turn him in from an objective or external perspective; given the whole of his beliefs, desires, and values, he has more reason *by his own lights* and from his own perspective. So in helping Jim escape, he does what he has most reason to do, even if his failure of deliberation prevents him from seeing this.

In freeing Jim, Huck displays an inconsistency between his beliefs and his actions, and this is one form of irrationality. But if he were to turn Jim in, he would be acting on the basis of deeply mistaken and unjustifiable beliefs, and these beliefs are *also* subject to rational criticism. Arpaly, Audi and McIntyre argue that this would represent an even more serious failure of rationality than the one he does display.

The general view that inverse akratics can be rational, and therefore praiseworthy, is compelling. Each of the existing particular arguments for this view, however, is flawed in some important way. The flaws in Audi and McIntyre's argument will be considered in Section III, below. What is worrying about Arpaly's argument is its heavy reliance on the notion of unconscious beliefs and desires. The appeal to the unconscious is unhelpful: the notion of "unconscious judgments" requires just as much explanation as "rational and

[34] Alison McIntyre, 'Is Akratic Action Always Irrational?' in Owen Flanagan and A.O. Rorty (eds.), *Identity, Character, Morality: Essays in Moral Psychology* (Cambridge, MA: MIT Press, 1990), 380-400., at 399.

praiseworthy akrasia", and so appealing to the former does not straightforwardly explain the latter. What is required is an explanation of structure of the inverse akratic's practical reasoning, which requires more than an appeal to unconscious beliefs.

To say that Huck's beliefs are "unconscious" might simply be to say that he has no occurrent thoughts about them while he acts, or that he has trouble summoning them as justifications after the fact. As we have seen, however, there is nothing particularly notable about either of these conditions. Inverse akrasia, however, *is* notable, and requires special explanation. Moreover, it even seems in principle possible for inverse akratics to act for reasons of which they are fully aware, but which they believe to be bad reasons, or at least insufficient reasons, for acting as they do. If such cases are possible, then an appeal to the unconscious will not explain them. The akratic agent might know *why* he acted: what makes his action akratic is not that his reason for action is unconscious, but that he does not judge it to be a sufficient reason for action, and yet he acts on it. Huck, for example, might have seen Miss Watson's ownership of Jim as providing an exclusionary reason, preventing what would otherwise be perfectly good reasons—friendship, loyalty, promises, and the rest—from carrying any deliberative force. In acting akratically, he would have acted on the basis of reasons he judged to be generally good ones, though in this case he had judged them to be without force. If so, an appeal to unconscious beliefs would not explain his actions.

How, then, to explain the deliberative outlook of inverse akrasia like Huck's? And how does that explanation leave room for praise? The first time Huck decides to turn Jim in, he is turned back by two things Jim says to him. As Huck paddles for the bank, Jim, unaware that Huck plans to turn him in, tells him that Huck is the best friend he has ever

had, and that he is now the *only* friend that Jim has. This is the remark that "takes the tuck" out of Huck. Nevertheless, he keeps paddling, though much slower, until Jim calls him "de ole true Huck, de on'y white genleman dat ever kep' his promise to ole Jim."[35] At this remark, Huck feels sick, and his resolve weakens so much that he turns back, despite his continued belief that he ought to turn Jim in. In fact, it is precisely the strength of that belief that leads Huck to feel sick in the face of his desire to help Jim escape.

It is Huck's emotional responses to Jim's appeals that lead him to act contrary to his beliefs. But the fact that it is these comments, in particular, that give rise to Huck's emotional reactions is significant. Why should an appeal to the value of friendship and the importance of loyalty and promises be described as, or give rise to what Bennett dismisses as "unreasoned emotional pulls"? Seeing someone as a friend, and understanding the value of friendship, involves a whole host of beliefs, judgments, and patterns of reasoning that are not properly described as "unreasoned."

Of course, Jim's appeal to friendship and promises prompts a series of emotional responses in Huck: love for his friend, shame at the thought of breaking a promise, perhaps fear for his friend's safety. But these responses are not unreasoned: they depend crucially on Huck's having particular beliefs about Jim, about friends, and about promises. They are, in other words, reasons-responsive in important ways. It is only if Huck believes in the value of friendship, believes that Jim is a friend, believes that promises are worth keeping, and believes that Jim is someone to whom promises can be made that Jim's comments can have any emotional effect on Huck. Otherwise, why

[35] Twain, *The Adventures of Huckleberry Finn*. Chapter 16.

would Jim's comments give rise to any emotional response at all, let alone weaken Huck's resolve?

Huck certainly acts on the basis of his emotional responses rather than his conscious beliefs about slavery. But this fact alone does not disqualify him from praise: we standardly praise people for their emotions and other attitudes, even though attitudes are generally not fully voluntary.[36] Moreover, Huck's emotions and attitudes are rationally grounded: they depend on his having a series of beliefs about Jim, friendship, promises, and loyalty. But though they are rationally grounded and connected to beliefs, Huck's emotions are not themselves beliefs. They have an evaluative and affective component that standard beliefs lack, and they are motivating in a way that beliefs are not.[37]

Huck's reasons for action, then, are his emotional responses to Jim's appeal: the pleasures he takes in friendship, the fear he feels of the potential suffering his friend faces, the shame he feels at the thought of breaking a promise. These affective responses are not, however, simply "unreasoned emotional pulls." They are rationally grounded. If Huck acts on the basis of emotional pulls, he nevertheless acts for reasons, and, in freeing

[36] For an excellent discussion of responsibility for attitudes, see Angela Smith, 'Responsibility for Attitudes: Activity and Passivity in Mental Life', *Ethics*, 115 (2005), 236-71. See also Raz, *Engaging Reason*.

[37] This account obviously presupposes a modest cognitivism about the emotions, according to which emotions have a cognitive or judgmental component. See, for example, Patricia Greenspan, *Emotions and Reasons: An Inquiry into Emotional Justification* (London: Routledge, 1988), Ronald De Sousa, 'Emotional Truth', *Proceedings of the Aristotelian Society: Supplementary Volume*, 76 (2002), 247-63. The account is modest because it only presumes that some emotional reactions, like Huck's, have an important cognitive component, and does not require that *all* emotions share this feature, or that judgments are *constitutive* of emotions. As such, it does not fall afoul of the sorts of objections to cognitivism raised in, for example, Jenefer Robinson, 'Startle', *The Journal of Philosophy*, 92 (1995), 53-74.

Jim because Jim is a friend to whom he has made a promise, Huck clearly acts for *good* reasons. To that extent at least, he is rational. Showing that his reasons are good ones, however, does not remove the accusation of akrasia, since after all, it appears that Huck does not see his reasons as good ones. So Huck might act for good reasons, but it remains to be seen if his deliberative outlook is praiseworthy, or if his akrasia disqualifies him from praise.

III. Comparing standard and inverse akrasia

Standard akrasia involves agents intentionally acting contrary to their own all-things-considered judgement. For example: an agent has a series of beliefs about the value of health and the unhealthy properties of crème brulé. On the basis of these beliefs, he forms the all-things-considered judgment that he ought to refrain from eating crème brulé. This judgment might be formed on the basis of conscious deliberation of his beliefs, but it need not be: it is enough that he would appeal to those beliefs in order to justify his judgment. In addition to his judgment that he should refrain, he also has a desire to eat crème brulé. This desire leads him to intentionally, but akratically, eat it, despite the persistence of his judgment.

In such cases, it is the agent's actions that are irrational and stand in need of an explanation. His actions are out of line with his all-things-considered judgment about how to act, and it is irrational, *from the agent's point of view*, to act contrary to this judgment. His desire does not necessarily stand apart from his judgment—he may well have taken the pleasure of crème brulé into account in making his judgment, and decided

that it was insufficient to outweigh the value of health. Nevertheless, he acts on his desire. In so doing, his action is clearly irrational.

Both Audi and McIntyre argue that, since there is a difference between what an agent believes he has most reason to do and what he in fact has most reason to do, it is possible for akratic agents to act rationally. This is a powerful argument, since it rejects a central assumption of the standard account of akrasia. But despite the force of their challenges, both concede too much to the standard view, since both accept that cases of inverse akrasia such as Huck's are properly described as acting contrary to judgment. For Audi, an akratic actions can be rational when it accords with an agent's "overall grounds of rationality better than does a practical judgment it contravenes."[38] This may be an accurate description of some cases of inverse akrasia, but it is not at all clear that inverse akratics like Huck actually *do* act contrary to their judgments. They certainly act contrary to their *beliefs* about how they ought to act, but there may be an important difference between what they believe and what they ultimately judge they have most reason to do.

Inverse akrasia is importantly different from the standard crème brulé case described above. If Huck were just like the standard akratic, we would say that, on the basis of his beliefs about slavery and property, he formed the all-things-considered judgment that he should turn Jim in. On the basis of a desire for Jim's freedom, however, he akratically frees Jim. If that were the correct description, Huck's behaviour would be

[38] Audi, 'Weakness of Will and Rational Action', at 279. McIntyre speaks in similar terms, saying for example that "akratic action may not be irrational if the agent is motivated by some consideration that could also have led her to revise the reasoning that led to her practical conclusion." Here, 'practical conclusion' carries the same meaning as 'practical judgment, and McIntyre is arguing that inverse akratic agents do *not* revise their conclusion prior to acting, despite having good reason to do so. McIntyre, 'Is Akratic Action Always Irrational?' at 390.

irrational in just the same way as the standard akratic's, and it would be puzzling to understand how he could merit praise for irrationally acting against his better judgment.

The standard description, however, gives far too much weight to Huck's conscious beliefs about slavery and property, and not nearly enough to his emotional evaluations of the situation. Huck certainly *believes* that slaves are property, that Jim is a slave, and that taking property is stealing. It is less clear, however, that Huck's all-things-considered judgment is that he should turn Jim in. Why should we associate Huck's judgment with his beliefs? After all, beliefs are just one intentional attitude among many. Huck's emotional evaluations of Jim, his affective responses to Jim's appeal, and his intentional actions in helping Jim escape all suggest that Huck *in fact* judges that he should help Jim escape rather than turn him in. This judgment might not fit with his beliefs, but that only shows that there is more to judgment than belief.

It is easy to simply assume that an agent's judgment are the result of her deliberations about the normative force of her various beliefs. This assumption is easy in part because of the deliberative implications of the phrase 'an all-things-considered judgment', which naturally suggests that all things were, in fact, considered and weighed. But we can and do act intentionally without engaging in such deliberation.[39] In fact, it seems at best misleading to say that Huck's 'all-things-considered judgment' was that he should turn Jim in. As Bennett points out, Huck failed to consciously consider *any* reasons for helping Jim escape when he was deliberating about what to do. Rather, he

[39] Both Audi and McIntyre primarily argue that akratics can act rationally even when the judgment they act against was the product of deliberation and reflection. This is no doubt true, and makes for a more powerful argument, but it also concedes (if only for the sake of argument) that deliberation and reflection are the norm for intentional actions. As I have argued, this concession should be resisted.

considered only his beliefs about slavery and property. It was only when he was faced with the prospect of actually turning Jim in that all of the considerations in favour of helping Jim entered his mind in any way. It was in light of these considerations that Huck acted, even though none of these considerations was entertained as a belief or entered into his conscious deliberations.

If this is right, then Huck does not act contrary to his all-things-considered judgment. Rather, like Lord Fawn, being faced with the actual prospect of turning Huck in gives rise to emotions that reveal his true judgment: he in fact judges that he should *not* turn Jim in, despite his beliefs to the contrary. Cases like Lord Fawn's make the potential difference between belief and judgment clear. To discover a moral incapacity in the way the Huck and Lord Fawn do is to discover that, whatever one *believed* one judged to be the best course of action, one was *mistaken* in that belief. That is why moral incapacities have the character of discovery as well as decision: they involve discovery that one's belief about one's judgment was mistaken.

Though belief and judgment are not the same intentional states, they are obviously closely connected. Beliefs serve as the basis for judgments, and judgments can be expressed as beliefs: in standard cases, someone who *judges* that she should do *x* will also *believe* that she should do *x*. Huck's judgment and (some of) his beliefs are in conflict, but his judgment cannot be wholly separated from his beliefs. If it is his decisive best judgment that he should help free Jim, he must have some beliefs that support and reflect this judgment. As indeed he does: Huck believes that Jim is his friend, that friendship and loyalty are important, and that promises should be kept. Huck would likely

assent to each of these beliefs, if asked. So his judgment that he should help Jim is not completely divorced from his beliefs, and indeed reflects many of those beliefs.

Of course, there is an important difference between Huck and Lord Fawn: the latter *endorses* his judgment after acting in a way that the former does not. In doing so he brings his beliefs, judgment, emotions, and actions into harmony. This harmony means that Lord Fawn is not akratic, despite acting contrary to what he had believed to be his judgment. Huck, however, continues to have mistaken beliefs about his judgment even after he acts. So his beliefs, judgement, emotions and action remain in conflict. This means that Huck continues to display an important failure of rationality that Lord Fawn does not.

Huck's irrationality is not that there is no rational connection between his beliefs and his judgment, since that would call into question his judgment as well as his beliefs. Rather, the source of his irrationality is that he does not focus his attention on the right beliefs. He cannot quite see how his beliefs about friendship are relevant, and he cannot bring himself to see how his beliefs about slavery and property are *not*. So even after he acts on his rational judgment that he should help Jim escape, he cannot *see* what he has done in this way. His beliefs about slavery get in the way. Similarly, he cannot see his beliefs about friendship and promises as relevant to Jim's freedom. As a result, Huck feels the intense mental anguish of regret: he cannot avoid the belief that he as acted immorally, since he cannot get rid of the belief that slaves are property. But this is a criticism of the rationality of Huck's system of beliefs, which he has not brought into order: it is not, in the first instance, a criticism of the rationality of his judgment or his

action. His beliefs about friendship and promises are connected to his emotional reaction to the situation, and this reaction is the source of his judgment.

Huck's failure is closely related to standard akrasia. If the argument I am making is correct, however, it is not the same. In standard cases, it is the akratic agent's *actions* that are irrational, since they are out of line with his all-things-considered judgment. In inverse cases like Huck's, however, it is the akratic agent's *beliefs* that are irrational, and not his actions, since it is his beliefs that are in conflict with his judgments. If this is right, then there is no problem in assigning praise to the actions of inverse akratic agents such as Huck on the grounds that their actions are irrational. Their actions are rational; it is their beliefs that are subject to criticism.[40]

This is an important difference: properly understood, inverse akratics do not act irrationally, since they do not act against their better judgment. Perhaps this difference is too important for both cases to be helpfully described as akratic. Inverse akratics are different in many ways from standard akratics, but there are nevertheless important similarities. Both are *subjectively* irrational in the same way. Both inverse and standard akratics have conflicts between belief, judgment, and action, and both *experience* their actions as irrational, as not reflecting their beliefs about what they should do. As a result, both experience the regret characteristic of akrasia. The difference is that inverse akratics

[40] Inverse akrasia is obviously related to akratic *belief*, or where an agent "believes against his better epistemic judgment." John Heil, 'Doxastic Incontinence', *Mind*, 93 (1984), 56-70. The two are not identical, however. For one thing, it is not clear that Huck "believes against his better epistemic judgment" that slaves are property —the judgment to which his belief is opposed is not, in the main, an epistemic one. Even if akratic belief is expanded to include non-epistemic judgments (as suggested in Mele, *Irrationality*. Chapter 8) it is not clear that it accurately describes Huck. Huck's problem is that he sees this belief as relevant to how he should act, not simply that he believes it.

have made a judgement about how they ought to act on the basis of their (rationally grounded) emotional reactions, and this judgment corresponds with their actions. From the inside, we might say, standard akrasia and inverse akrasia are difficult to distinguish. It is only from the outside that we can assess the connection between the agent's judgments, beliefs, and actions, in order to determine whether it is the agent's *action* or his *beliefs* that are out of line and are therefore irrational.

One upshot of this argument is that, when it comes to the assessment of someone's moral deliberative outlook, beliefs are not the only relevant consideration. We sometimes get a much better handle on the rationality of a moral outlook by considering emotional reactions, and the judgments that embody from them, than we do by considering moral beliefs. Moral knowledge, in other words, is not exclusively located in moral beliefs, but is potentially spread throughout beliefs, emotions, evaluations, desires, judgments, intentions, and other intentional states.

There is something potentially liberating about this claim, since it means correctness of moral belief is not a precondition of a rationally defensible moral outlook. In fact, one's moral deliberative outlook can be, like Huck's, praiseworthy in many respects despite the presence of deeply mistaken beliefs. It also means that disputes about the correctness of moral beliefs do not need to entail disputes about he correctness of a moral outlook more broadly. Similar emotions and judgments can be compatible with a wide range of moral beliefs. This means that the devoutly religious and the firmly secular can share many moral judgments even if they do *not* share many moral beliefs.

Despite having deeply mistaken beliefs, and despite a form of practical irrationality similar to akrasia, Huck, and inverse akratics like him, acts rationally.

Moreover, Huck's action are praiseworthy, Not only does he act for the right reason: there are other aspects of the structure of his deliberative outlook that make him an appropriate candidate for praise. His acting for the right reason is associated with the appropriate emotional responses: a feeling of love for his friend, a recognition that turning Jim in is 'unthinkable', and even shame and disgust at the idea of betrayal. Of course, Huck is particularly lacking in self-awareness, and so we can recognize what he cannot: that the shame he feels when he does not turn Jim in due more to the thought of turning in a friend as it is to dismay at his own weakness. Huck, in other words, lacks a considerable degree of self-knowledge: he does not fully appreciate his own judgments or the reasons for which he acts, and he lacks conscious appreciation of the source of his own emotions, even as those emotions express his genuine moral judgments. Despite this lack of self-knowledge, however, Huck's deliberative outlook merits a significant degree of praise: he acts for morally worthy reasons, and, in so doing, he has a series emotional responses that express a morally praiseworthy appreciation of the situation.

It is important to recognize, however, that this does not mean that Huck would not be *more* rational and *more* praiseworthy if his beliefs, emotions, judgements, and actions were in harmony. If Huck had greater self-knowledge, and a better appreciation of both the reasons for which he acts and the source of his emotions, then Huck's deliberative outlook would be even more morally worthy.

It might be objected that Huck's psychic harmony is irrelevant to his moral worth. It would certainly be better *for Huck* if his beliefs, emotions, judgments, intentions, actions, and so on were all aligned, but why should we suppose that it is relevant to our level of praise?

The answer, in part, is that Huck's actions, while minimally rational, would have been more rational with a tighter connection between his beliefs and his judgment, and a greater awareness of the reasons for which he acted. Huck's judgment that he should help Jim is based in part on the beliefs that ground his emotions: his belief that Jim is a friend, that friends should be helped, that promises should be kept, and so on. But Huck's judgment, while praiseworthy, is still contingent in worrisome ways. First, since he lacks an appreciation of the reasons for which he acted (and so the source of his emotional responses), the connection between these reasons and his action is not as close as it should be. It is the gap between Huck's actual reasons and his understanding of them that makes him akratic, and also makes his actions worryingly contingent: he could have easily acted on his mistaken understanding of his reasons instead. So a greater degree of self-knowledge would have made Huck's actions considerable less contingent, and so would have made his deliberative outlook more stable and hence more praiseworthy.

Second, though friendship and promises are good reasons for helping Jim, the fact that they are Huck's only reasons means that he would have likely felt little hesitation at turning in a stranger. Huck would be even *more* praiseworthy if his judgment was based, not simply on the thought that *Jim* ought to be freed, but that *all* slaves deserved the same. And it is precisely his mistaken beliefs that block him from being able to make this judgement. So while Huck acts for good reasons, his lack of psychic harmony means that he could be even more praiseworthy than he already is.

Inverse akratics such as Huck are certainly less than perfectly rational, but then so are most of us, most of the time. It is possible to be *less that perfectly rational* without being completely *irrational*. Though Huck would of course be a better person if he were

equipped with all of the right beliefs about the wrongness of slavery (and none of the wrong ones). No doubt Huck would also be more deserving of praise if he had acted from a firm and unchanging disposition, in full knowledge that saving Jim was the right thing to do, with pleasure, and *because* saving Jim was the right thing to do. But if the only way our actions can be rational and, in moral contexts, merit praise, is if we have full knowledge, all the relevant belief, no false beliefs, then few of us ever manage to be either rational or praiseworthy. To be less than fully rational is not the same as being irrational, and to have a less than fully praiseworthy deliberative outlook does not mean that one's deliberative outlook is not praiseworthy *at all*. Likewise, to be less than fully virtuous is not the same as being vicious. Imperfect agents, and even some akratics, can be rational and merit praise despite their imperfections. There is, then, no principled asymmetry between praise and blame in the case of akrasia. Though inverse akratics still display a failure of rationality, and though they are in some ways very different from standard akratics, they can and do merit praise as well as blame.

Chapter 4: What Is Wrong With Hypocrisy?

Hypocrisy implies a moral code
George Orwell

Hypocrisy is interesting in part because it is just so *prevalent*: our culture, at least, seems to be swimming in it, and each of us contributes our share. And almost as prevalent as hypocrisy is condemnations of it: it is, in Judith Shklar's words, "the only unforgivable sin', inexcusable even for those who can justify almost any other vice."[1] Hypocrisy's cast of characters is a wide-ranging and diverse group. Some present a public façade of virtue that hides a private life of vice; others act in morally desirable ways, but appear to do so for the wrong sorts of reasons; still others claim to have the noblest of values, but never seem to act on them. An adequate account of hypocrisy should do at least three things: it should capture its *scope*, to make sense of the wide cast of characters that exhibit this common, yet puzzling vice. Second, it should explain how hypocrisy is psychologically possible. Finally, it should explain why hypocrisy arouses such moral outrage. After all, there is something puzzling about hypocrisy's pervasive condemnation. Hypocrisy is often taken to involve an assumption of virtue that is largely pretence, motivated by a desire to *appear* virtuous rather than to actually *be* virtuous. Yet often, this means that hypocrites take pains to act similarly to virtuous agents. They may do so for the wrong reasons, and so may not deserve our praise, but it seems preferable to a complete disregard for virtue. Why, then, are we so hard on hypocrites?

It is useful to begin the discussion with a sample hypocrite. Few writers have

[1] Judith Shklar, 'Let Us Not Be Hypocritical', *Ordinary Vices* (Cambridge, MA: Belknap Press of Harvard University Press, 1984), 45-86 at 45.

sought to expose hypocrisy more clearly and with greater intensity than George Orwell. Though he was a committed Socialist—he fought as a volunteer in the Spanish Civil War—Orwell was nevertheless deeply critical of what he saw as the intellectual dishonesty of much left-wing politics.

> All left-wing parties in highly industrialized countries are at bottom a sham, because they make it their business to fight against something they do not really wish to destroy. They have international aims, and at the same time they struggle to keep up a standard of life with which those aims are incompatible. We all live by robbing Asiatic coolies, and those of us who are enlightened all maintain that those coolies ought to be set free; but our standard of living, and hence our 'enlightenment', demands that the robbery shall continue. A humanitarian is always a hypocrite.[2]

If humanitarians are always hypocrites, how should we explain their hypocrisy? The few philosophers who have developed accounts of hypocrisy generally agree that it is characterized by a particular sort of deception: according to this standard account, hypocrites *pretend* to be morally better than they really are. On this view, Orwell's humanitarian does not *really* think that ending colonialism is important: he just pretends to because it is in style, or so that others will admire him.

In my view, this deception account of hypocrisy is a complete failure. Either it completely fails to capture the diverse cast of characters that we rightfully accuse of hypocrisy, or else, by appealing to *self*-deception to capture hypocrisy's scope, it depends on a mistaken analogy between self-deception and interpersonal deception and fails to properly explain just what is morally objectionable about the vice.

I offer a different account of hypocrisy, one that dispenses with the emphasis on

---

[2] George Orwell, 'Rudyard Kipling', *The Collected Essays, Journalism, and Letters of George Orwell, Volume 2: My Country Right or Left* (Harmondsworth: Penguin, 1970), 215-29 at 218.

deception while allowing that many hypocrites can be self-deceived. On my view, hypocrites all care too much about their image for having certain values, and this excessive concern for their image leads them, in many different ways, to fail to honour the values that they claim to have. But their failure need not involve deception of either themselves or others, though it can often involve *self*-deception and related failures of self-knowledge. In rejecting the emphasis on deception, I urge that we reconsider the moral status of hypocrisy. Many hypocrites do exhibit an important moral failure, but their failures are often relatively minor, and are far from being "the only unforgiveable sin." In fact, I argue that hypocrites can even be praiseworthy in actions that express the values that they hypocritically hold.

I. The deception account of hypocrisy

According to the most common account, what all cases of hypocrisy have in common is a particular kind of deception. The hypocrite is a liar who lies about *herself*: she aims to deceive others about her true beliefs, desires, and intentions—in short, about her intentional attitudes.[3] As Eva Feder Kittay puts it, the hypocrite "pretends to be better than she is, given a norm or set of expectations within a domain in which sincerity really matters."[4] This helps explain why hypocrisy is seen as an important moral failure: the

[3] Of course, not all self-referential deception is hypocritical, since actors are not hypocrites as we have come to understand the concept. In part, this is because even the most skilful actors do not genuinely deceive their audience, who merely agree to 'suspend disbelief.' But the Greek *hypokrisis*, from which 'hypocrisy' derives, can simply mean 'playing a part,' and carries no negative connotation. Actors are not hypocrites in the pejorative sense, but the history of the concept of hypocrisy does involve the notion of acting or pretending.

[4] Eva Feder Kittay, 'On Hypocrisy', *Metaphilosophy*, 13 (1982), 277-89 at 277.A similar

hypocrite often aims to get the benefits of being moral without paying the necessary costs.[5]

It is important for the deception account that the hypocrite's pretence is intentional: she actually *aims* at deception. Christine MacKinnon has advanced the most forceful statement of this view: the hypocrite "must intend some harm—if only that of deceiving or manipulating her fellow agents", and she "must be aware that she is manipulating the judgments of her audience."[6] The deception of the hypocrite cannot be accidental—she must know that her actions will create a false impression about her, and do those actions *for that reason*.

The standard example of this deceptive form of hypocrisy is Molière's eponymous character Tartuffe, who pretends to be extremely pious only so that he can take advantage of his legitimately religious host, Orgon. Tartuffe's hypocrisy is clear-eyed: he is not pious in the least, he knows full well that appearing pious is to his advantage, and he intentionally acts piously so as to procure that advantage. On the deception account, hypocrites merely pretend, like Tartuffe, to have beliefs, desires, and

---

account is advanced by Bela Szbados and Eldon Soifer, who argue "it is not the pretence of self-interest alone that is crucial for hypocrisy, but rather the use of deception to gain an unmerited self-interested reward." Bela Szabados and Eldon Soifer, *Hypocrisy: Ethical Investigations* (Peterborough: Broadview, 2004) at 166.

[5] It is worth noting that, though hypocrisy attracts a great deal of moral censure, the values at issue need not be specifically moral. One of the examples I will consider is of someone who has a hypocritical regard for aesthetic values, and it may even be possible for a hypocrite to care too much about his image for having *immoral* values. Feder Kittay's general definition captures this well: the hypocrite pretends to be better "given a norm given a norm or set of expectations within a domain in which sincerity really matters." Feder Kittay, 'On Hypocrisy', at 277. On my account, hypocrisy can be a failure of practical or evaluative reason in general, rather than a specifically moral failure.

[6]Christine Mackinnon, 'Hypocrisy, with a Note on Integrity', *American Philosophical Quarterly*, 28/4 (1991), 321-30 at 322-3.

motivations that they do not have with the deliberate aim of creating the impression of being morally better than they in fact are.

i. A gap between intention and action?

Those who defend the deception account of hypocrisy explain the hypocrite's psychology by claiming that hypocrisy opens a gap between intention and action: while agents standardly act in ways that express their intentions, hypocrites do not, and instead act in ways that are opposed to their intentions.[7] It is this gap that creates the deception that characterizes hypocrisy: since, according to the standard view, the hypocrite's acts are severed from his intentions, it is impossible to read off his intentions and motivations from his actions.

It is difficult to reconcile the standard account's emphasis on intentional deception with the claim that the hypocrite's action is severed from his intention, such that his actions do not express those intentions. The problem is that this claim makes the hypocrite appear akratic, somehow failing to act as he intends. But hypocrites like Tartuffe are not akratic. Quite the contrary: Tartuffe knows precisely what he wants to do—create the impression of piety—and his actions are chosen with that end in mind.[8] There is no more of a gap between what Tartuffe intends to do and what he in fact does

[7] For example: The hypocrite ""ruptures the link between desires, reasons for action, and actions" Christine Mackinnon, 'Hypocrisy and the Good of Character Possession', *Dialogue*, XLI (2002), 715-39 at 735; he "severs the act from the intention" Feder Kittay, 'On Hypocrisy', at 285; and "by divorcing motive from action the hypocrite distorts the psychology of action" Mackinnon, 'Hypocrisy, with a Note in Integrity', at 327.

[8] We might say that the gap in hypocrisy is *external*, between the hypocrite's actual intentions and the intentions others are likely to attribute to him. The gap in akrasia, on the other hand, is *internal*, between the agent's own judgments and his actions.

than there is in the case of the person of integrity. When MacKinnon says that the hypocrite "distorts the psychology of action", she cannot mean that, in acting, the psychology of the *hypocrite* is dramatically unlike that of the standard agent. Rather, she must mean that our *psychologizing* about him is unreliable, since we are unable to discern his motives. If the hypocrite is the sort of intentional deceiver that the standard account describes, then the real gap is not between the hypocrite's intentions and his actions, but between his actual intentions and the intentions that others are liable to attribute to him on the basis of his actions.

Such a view would characterize hypocrisy as a gap between what the hypocrite intends and what others *believe* the hypocrite intends. Not all such gaps are evidence of hypocrisy, however, since it is possible to be mistaken about the intentions (as well as the other intentional attitudes) of people who are not hypocrites. This is for at least two reasons. First, while we can generally infer the attitudes of others from their behaviour (including their speech), such inferences are not infallible, since none of us are completely transparent. In fact, the very possibility of hypocrisy *depends* on our lack of transparency. Even for the most informed and impartial observer, the attitudes of others are underdetermined by the available evidence.[9] Second, the problem of underdetermined evidence is compounded by the fact that very few of us ever count as perfectly impartial and informed observers. The stories we tell to explain the actions of others can say as much about *us* as they do about those we are looking to explain.[10] Our assumption that

[9] As Anscombe famously points out, a single physical act can be the expression of any number of different intentions. *Intention* at §23.

[10] Nomy Arpaly points out, that we do not, as Davidson supposes, always accept the most rationalizing explanation of the behaviour of others. Often, "our default assumption when

others are like us may generally be warranted, but it can also cause us to radically misinterpret the actions, intentions, and motives of those who have very different beliefs, desires, values, and characters.

There are therefore two barriers to a clear understanding of the motives and intentions of others: we frequently do not have enough evidence to be able to decisively determine someone's intentions, and we frequently make poor use of the evidence that we do have. And, even if we make flawless use of the best available evidence, we may still fall into error. So unless we are to count a person as a hypocrite every time someone makes an unwarranted assumption about her intentions, not all examples of a gap between actual and perceived intentions or motives are cases of hypocrisy. An actual gap between real and perceived motivation is therefore not sufficient to characterize hypocrisy. Nor is an actual gap necessary—much hypocrisy is obvious to us, and the failure of the hypocrite to mislead his audience as to his real motivation does not make him any less hypocritical. On the deception account, then, he is a hypocrite not because there is an *actual* gap between real and perceived motivation, but because he intends to deceive others about his real motivation.

The presence of an actual gap between intention and action, or between real and perceived motive, is therefore neither necessary nor sufficient for hypocrisy. The common claim that the hypocrite "divorces act from intention", which looked like an

---

trying to understand a person is that the person is *similar to us.*" Most of us are well aware that we are not always rational—we suffer from akrasia, self-deception, irrational fears, and so on. So we will often assume that others suffer from these faults as well. And the more flawed we are, the more likely we may be to attribute flaws to others. Nomy Arpaly, *Merit, Meaning, and Human Bondage: An Essay on Free Will* (Princeton: Princeton University Press, 2006) at 72.

explanation of the hypocrite's distinctive psychology, is in fact simply a restatement of the claim that hypocrites aim to deceive others about their intentional attitudes. Our psychologizing about hypocrites may be unreliable, but, if we accept the deception account, the hypocrite's psychology itself is no different from that of the standard liar.

ii. The deception account and the wrongness of hypocrisy

The deception account deals with the blame criterion in a straightforward way. Since hypocrisy involves intentional deception, it so wrong for the same reasons that lying is wrong. The deceptive hypocrite, like the liar, manipulates others for his own ends in ways that he would be unable to justify to them.[11]

This sort of unjustifiable manipulation of others would appear bad enough, but those who advance the deception account of hypocrisy seem to think that it does not fully capture the evil of hypocrisy. It is not merely that hypocrites manipulate others, they say, but that they actively aim to subvert morality by trying to gain the advantages of a good moral reputation without appreciating the value of morality. In deceiving others about his true attitudes, and, more deeply, abut himself, the deception account portrays the hypocrite as "undermining morality and exhibiting a disdain for its practitioners."[12] This makes the hypocrite a very nasty character indeed, worse even than the avowed egoist

[11] For a more extended account of the wrongness of lying along these general lines, see T.M. Scanlon, *What We Owe to Each Other* at 317-22.

[12] MacKinnon, 'Hypocrisy, with a Note in Integrity', at 323. Feder Kittay expressed the same point in almost identical language: the hypocrite "undermines the very conception of that to which he pretends, be it piety, virtue, or friendship." Feder Kittay, 'On Hypocrisy', at 286. MacKinnon later argues that the hypocrite undermines morality because he "undermines our collective practices of judging persons." MacKinnon, 'Hypocrisy and the Good of Character Possession', at 716.

who does not care for morality.[13] No wonder, then, that hypocrisy is held to be so unforgivable.

This emphasis on the extreme harms of hypocrisy is exaggerated. The idea seems to be that the merely vicious only wrong *others,* while hypocrites harm both others and *morality itself*. It is hard to understand just what this claim is supposed to amount to. It cannot be that hypocrisy undermines morality by showing others that it is possible to succeed without following the rules, since hypocrisy is meant to be undetected: if morality is undermined by bad examples, hypocrites are far *better* than those who openly flout the rules, since hypocrites at least *try* to hide their vice. If hypocrisy is "the compliment that vice pays to virtue", then we might reasonably suppose that it is better to pay morality a compliment than to openly deliver it an insult. Nor can it simply be that hypocrisy, alone among crimes, depends for its success on the presumption that others will follow the rules. Indeed, for those with Kantian leanings, *all* wrongdoing will share this feature to some extent. If hypocrisy undermines morality in this way, it is not unique in doing so.

More importantly, the wrong of manipulating others is not best understood as a wrong committed against morality itself. This is something Kant emphasizes when he says that "a lie always harms another; if not some other human being, then it nevertheless does harm to humanity in general, inasmuch as it vitiates the very source of right."[14] In doing harm to 'humanity in general,' lying wrongs *all of us,* rather than wronging an

---

[13] Soifer and Szabados assert that "intuition strongly suggests" that hypocrites are even worse than "acknowledged egoists." Eldon Soifer and Bela Szabadon, 'Hypocrisy and Consequentialism', *Utilitas* 10 (1998), 168-94 at 172.

[14] Immanuel Kant, 'On a Supposed Right to Lie Because of Philanthropic Concerns', *Groundwork for the Metaphysics of Morals* at 65.

abstract entity such as 'morality.' A crucial insight of the Categorical Imperative is that, in acting in ways that cannot be universalized, wrongdoers act in ways that they cannot justify to others.[15] This is wrong because it involves a failure to treat others as autonomous agents, worthy of respect, not because it involves a harm to morality as such, whatever that might mean.

Failing to treat others as autonomous agents is precisely what the hypocrite does when he seeks to manipulate others through deception. MacKinnon and Feder Kittay's emphasis on the harms hypocrisy does to morality is a distraction from what is genuinely objectionable, which is that, on their view, hypocrites intentionally manipulate others in order to gain an unjustifiable advantage.

II. Non-lying hypocrites

The deception account's emphasis on intentional deception makes the psychology of the hypocrite easy to understand, and ties it directly to the wrongness of hypocrisy. It therefore elegantly unifies two of the three criteria for an account of hypocrisy. The failure of the account as it stands is its scope. Tartuffe, the clear-eyed liar, is one sort of hypocrite, but if we take seriously the way in which the concept of hypocrisy is deployed, he is far from the only member of the cast. Quite often, we accuse others of hypocrisy without meaning to accuse them of *lying* about their attitudes—what makes Orwell's humanitarian a hypocrite is a deep lack of self-awareness, rather than a calculated attempt

[15] This is part of the demand expressed in the third formulation of the Categorical Imperative: "act in such a way that you treat humanity… always at the same time as an end, never merely as a means." Immanuel Kant, *Groundwork for the Metaphysics of Morals* at 36.

at deception. Orwell is not suggesting that left-wing politics is a sham because it is an elaborate *lie*, but because, deep down, 'enlightened' people are not actually prepared to act on the principles that they endorse. Humanitarians may be hypocrites, but if so, it is not because they are, like Tartuffe, devious schemers.

If Orwell is right, and his socialist colleagues are hypocrites, then not all accusations of hypocrisy are meant to suggest that the target is actually aiming to deceive about his attitudes. We can all recognize examples of this sort: the environmentalist who continues to drive to work in a car bearing a "Save the Planet" bumper sticker when she could take public transit is at this point a stock character in our political morality play. But though we accuse her of hypocrisy, it would be wrong to say that she was simply *lying* about her commitment to the environment. Her psychology is not that simple. Giving an adequate response to the scope criterion therefore requires an improved response to the psychological criterion as well.

i. Extending the deception account: self-deception?

Some defenders of the deception account recognize the problem it has with the scope of hypocrisy, and they try to extend the account by including the possibility of *self*-deception. Self-deception involves motivated false belief: believing something that is false because one wants it to be true. So perhaps the hypocritical humanitarian wants so much to convince others that he is better than he is that he has actually managed to convince *himself* that he is, in fact, better than he is. He will therefore resist the accusation of hypocrisy, not because he does not want to be found out, but because he does not believe it to be accurate. Instead of deceiving *others* about his attitudes, he will

have deceived *himself*. If successful, this appeal to self-deception will preserve the flavour of the deception account while accommodating the full range of hypocrites.

An appeal to self-deception in order to address concerns about the scope of their account is common for those who embrace the deception account of hypocrisy.[16] Even Christine MacKinnon, who initially presented an extremely narrow account of hypocrisy that made no room for self-deception, later modified her account to acknowledge that some hypocrites are "quite unself-conscious about the extent to which they misrepresent their real reasons for acting."[17] MacKinnon therefore allows that some hypocrites fall somewhat short of calculated lying, but she continues to insist that hypocrisy is characterized by a misrepresentation of the hypocrite's motives and reasons for action, and that this misrepresentation be carried out because the hypocrite wants to appear to be better than she in fact is.

It is certainly true that many hypocrites are self-deceived in various ways. If self-deception involves motivated false belief in the face of the evidence, then the hypocritical environmentalist who drives to work when she could easily take public transit, and yet persists in her beliefs she does all she can for the environment is self-deceived. After all, she wants to believe that she is environmentally pure, and her desire leads her to believe

---

[16] Szabados and Soifer who believe that deception is essential to hypocrisy, appeal to self-deception in exactly this way. They argue that self-deception explains "the existence of cases in which people are genuinely surprised to learnt that they have been hypocritical, without thereby conceding that there can be cases of hypocrisy that do not involve deception." Szabados and Soifer, *Hypocrisy: Ethical Investigations* at 256. Daniel Statman goes further, and argues that self-deception and hypocrisy cannot be clearly distinguished.Daniel Statman, 'Hypocrisy and Self-Deception', *Philosophical Psychology,* 10 (1997), 57-78.

[17] Mackinnon, 'Hypocrisy and the Good of Character Possession', at 719. The narrower account is found in her Mackinnon, 'Hypocrisy, with a Note in Integrity.'

it in the face of the evidence. Moreover, she is deceived about the extent to which her actions reflect her self-image, and so she is self-deceived about *herself*. Nevertheless, I will argue that an appeal to self-deception cannot save the deception account of hypocrisy by extending its scope.

The heart of the problem with extending the deception account is that it relies on an analogy between self-deception and interpersonal deception. This analogy is important because it allows the deception account to argue that clear-eyed hypocrisy (like Tartuffe's) and self-deceptive hypocrisy (like Orwell's humanitarian) are essentially similar, and because they are similar, they are wrong for the same reasons. If lying and self-deception are *not* relevantly similar, then what looks like an elegantly unifying analysis of seemingly diverse phenomena is reduced to a description that simply lists two very different conditions.

In fact, the analogy between interpersonal deception and self-deception breaks down in two important ways. The first breakdown is that interpersonal deception is intentional, a point exploited by those who appeal to self-deception. So Daniel Statman argues that "the best way for us to create a certain image in the eyes of others is to believe this image ourselves. For their deception to be most effective, hypocrites should believe their deception, i.e. be [self-deceived.]"[18] Statman clearly sees self-deception here as part of an intentional strategy for deceiving others. The problem with this strategy is that the best account of self-deception holds that it is *not* intentional.

It is true that some explanations of self-deception claim that it can be

[18] Statman, 'Hypocrisy and Self-Deception.'

intentional.[19] Such accounts all have to explain away the paradoxical idea that the self-deceived agent takes his belief that p is false as his reason for believing that p is true, and thus manages to believe both p and ~p.[20] The solutions to this paradox are unsatisfying, since they depend on a problematic homuncularism or to ad hoc divisions within the mind that leave unexplained exactly how self-deception can ever proceed.[21]

The paradoxes of self-deception only arise if it is understood as intentional. Deflationist accounts of self-deception argue that, once we recognize that our desires can bias the ways that we gather and interpret evidence without our ever *intending* to form false beliefs, the paradoxes dissolve, and with them the need for controversial homuncular solutions.[22] But if we do not mean to deceive *ourselves*, then self-deception cannot be part of an intentional strategy for deceiving *others*. So self-deception cannot play the same role that interpersonal deception played in the narrow version of the deception account. As I will argue below, this threatens the deception account's explanation of the wrongness of hypocrisy.

The second breakdown in the analogy between interpersonal and self-deception is that intentionally deceptive hypocrites deceive others about their attitudes in a way that is

---

[19] See especially Donald Davidson, 'Deception and Division', *Problems of Rationality* (Oxford: Clarendon Press, 2004), 199-212, and David Pears, *Motivated Irrationality* (Oxford: Oxford University Press, 1984).

[20] The self-deceived person has to believe these things, in some sense, *at the same time*, since a) believing that p at T1 and that ~p at T2 can simply represent a change of mind rather than self-deception, and b) the self-deceived person's belief that p in some sense motivates and *sustains* the belief that ~p.

[21] Mark Johnston, 'Self-Deception and the Nature of Mind', in Brian Mclaughlin and A.O. Rorty (eds.), *Perspectives on Self-Deception* (Berkeley: University of California Press, 1988), 63-91, gives a decisive argument against such strategies.

[22] For an explanation of just how desires can lead to motivationally biased beliefs without the self-deceived agent's ever intending to adopt a false belief, see especially ibid., and Alfred Mele, *Self-Deception Unmasked* (Princeton: Princeton University Press, 2000).

simply not open to the self-deceived. For the deception account, hypocrites are distinguished from (mere) liars in that hypocrites deceive others about what they believe, desire, and value; in short, about the content of their intentional attitudes. If self-deception can extend the deception account, it must be possible for non-lying hypocrites to be self-deceived about the content of those own attitudes. But the analogy breaks down here as well. It is of course perfectly possible to deceive *others* with respect to the content of one's own attitudes. As I shall argue at greater length in Chapter 5, however, it is not generally possible to deceive *oneself* about the content of one's own attitudes. This disanalogy reflects the asymmetry between the first- and third-person perspectives on the content of our own attitudes: while others form beliefs about one's attitudes on the basis the evidence of one's behaviour, one's beliefs about one's *own* attitudes are not formed on the basis of inference from evidence (and certainly not from the evidence of our own behaviour) but are known, in Anscombe's phrase, "without observation."[23] Though there is considerable debate about the exact nature of self-deception, all parties to the debate agree that it is possible because our desires can bias the ways in which we gather and interpret evidence. We therefore cannot be self-deceived about the content of our own attitudes, because the mechanisms of self-deception can get no purchase on our beliefs about those attitudes.

The claim that we cannot be self-deceived about the content of our own attitudes is independent of the claim that self-deception cannot be intentional, since both the intentionalist and deflationist accounts appeal to similar evidence-biasing mechanisms in order to explain how self-deception occurs. The disagreement is about whether those

[23] Anscombe, *Intention* at §8.

mechanisms can operate intentionally. This means that an account of hypocrisy that appeals the analogy between self- and other-deception breaks down in two distinct ways—lying is intentional, while self-deception is not, and hypocrites can lie about the content of their attitudes, but we cannot be self-deceived about the content of those attitudes.

These two disanalogies between the psychology of lying and self-deception lead to a third problem with the attempt to extend the deception account by appeal to self-deception. Hypocrisy is held to wrong because it involves the selfish and unjustifiable manipulation of others. This assumes that the manipulation is *intentional*. But self-deception is not intentional, and so it cannot figure as part of a strategy for the manipulation of others. Once the deception account admits that hypocrites can be self-deceived, it loses its grip on what it is that makes hypocrisy morally objectionable, since it can say nothing about the blameworthiness of self-deceived hypocrites. But the deception account aims in large part to explain just *why* hypocrisy is so morally objectionable: this is why it emphasizes the analogy with lying. Rather than extending the scope of the deception account, the appeal to self-deception shows how the account is fundamentally misguided. Another approach is required.

III. Hypocrisy as misdirected concern

The deception account fails, since its attempt to meet the scope criterion both relies on a false analogy between self-deception and interpersonal deception, and invalidates its answer to the blame criterion. A plausible explanation of hypocrisy must begin by rejecting the single-minded emphasis on deception.

Nevertheless, the deception account began with an important insight: hypocrites want to appear better than they actually are. [24] They have a desire for an image of virtue that is out of line with their actual level of virtue. The deception account wrongly emphasized the role of deception in satisfying this desire, but was right that the desire itself is a key component of hypocrisy. On the account I will defend, hypocrites care too much about their image as people with certain virtues and values, and this excessive concern leads them to fail, in a variety of ways, to properly honour the values they claim to hold. In a phrase: hypocrisy is a failure to properly 'value one's values', brought on by an excessive concern for one's image as someone with certain values.

The claim that hypocrites have a desire to appear better than they are is ambiguous: it can mean that hypocrites know that they are not virtuous but simply aim to appear so, or—more broadly—it can mean simply that hypocrites aim to appear better than they are, without the condition that they *know* that they are doing so. The narrow deception account chose the first interpretation, and ran into the problem of scope. The account I defend, like the self-deception version of the deception account, prefers the second interpretation. This means that some hypocrites can display an important failure of self-knowledge: they do not know that they present themselves as better than they are. I will return to the question of the hypocrite's lack of self-knowledge below, after explaining just what makes the desire to appear better than one is hypocritical.

Though this desire is an important feature of hypocrisy, there is nothing *necessarily* hypocritical about it. Someone might have a very strong desire that others

[24] Feder Kittay says that the hypocrite "pretends to be better than she really is" Feder Kittay, 'On Hypocrisy', at 277.; Soifer and Szabados says he aims to gain "an unmerited self-interested reward." Szabados and Soifer, *Hypocrisy: Ethical Investigations* at 166.

believe she is better than she actually is, but only because she actually wants to *be* better than she is. The disappointment of those whose opinions we value can be a tremendous motivator. A desire for a better reputation than we deserve is not necessarily hypocritical if it is tied to an even stronger desire for self-improvement. The question is what must be added to the desire to make it hypocritical.

The deception account answers this question by emphasizing the necessary role of deception in satisfying such desires. Hypocrites aim to satisfy their desires by tricking others (or perhaps themselves) rather than by actually becoming more virtuous. As we have seen, this response is unpersuasive. There must be something else, other than deception, which can make this desire hypocritical.

The hypocrite's primary desire is for an *image* of virtue. The problem with this desire is not that it is it is essentially *deceptive*, but that it is *misdirected.* There is nothing hypocritical about wanting to be virtuous; quite the opposite, in fact. Nor need there be something hypocritical about the person who values her image as virtuous: perhaps what he most wants is to be a virtuous person, and he wants other people to rightly see him as such. Aristotle, after all, reserves his highest praise for the magnanimous person, who "thinks himself worthy of great things, and is really worthy of them."[25] The magnanimous person certainly cares for his image as someone who is virtuous, but that is because he truly is virtuous, and he rightly believes that virtue is worthy of honour, "the greatest of external goods."[26] As great as honour is, however, virtue is greater: the magnanimous person wants others to think he is virtuous because he truly cares about the

[25] Aristotle, *Nicomachean Ethics* at 1123b3.
[26] Ibid. at 1123b20.

virtue. That is why he only cares about being respected and honoured by other virtuous people. Honour, for him, is "an adornment of the virtues."[27]

So while there is nothing necessarily hypocritical about wanting others to recognize one's virtue, there nevertheless is, as Bernard Williams points out, something "suspect" about the person whose deliberations are guided by the question of whether his actions are virtuous. Kind and courageous people do kind and courageous things, but they rarely do them under those descriptions. When the kind person acts, her reasons are considerations such as "he is in pain", and "he needs it more than I do," and not, primarily, reasons such as "this would be kind and win me honour." This is why honour is an adornment to virtue. To think of one's own actions in terms of the virtues, says Williams, represents "a misdirection of the ethical attention," even though thinking of the actions of *others* in such terms is appropriate.[28] Such ethical attention is misdirected because it primarily considers how others would describe your actions, rather than how you ought to act. This sort of deliberation externalizes your moral attention, focusing it on the wrong sorts of considerations.

An externalized ethical attention is not necessarily *deceptive*, but it is misdirected, and such misdirected ethical attention is at the heart of hypocrisy. The hypocrite's attention is misdirected not simply because he has a desire to appear—to himself and to others—to have certain values and virtues, but because he cares *more* about this appearance than he does about actually having those virtues and values. In other words, he cares too much about his image as someone with certain values and not enough about

[27] Ibid. at 11245a1.
[28] Bernard Williams, *Ethics and the Limits of Philosophy* (Cambridge, MA: Harvard University Press, 1985) at 10.

the values themselves. Moreover, because he cares too much about his image, he fails, in various ways, to respect the values that he claims to hold. This is part of Williams' point: those who truly are virtuous do not think of their actions in terms of the virtues. This does not mean that they are *unaware* that their actions are open to such descriptions, but only that those descriptions do not play an active role in their deliberations about how to act.

This definition of hypocrisy involves an essentially comparative claim: a hypocrite's desire to have virtuous image is *stronger* than his desire to actually be virtuous. Hypocrisy does not simply involve caring about one's image: I just argued that it is possible to care about one's image without being a hypocrite. Nor are those who genuinely do care about the values they profess always innocent of hypocrisy. Hypocrites like Tartuffe, it is true, do not care at all about the values they profess, but Orwell's humanitarian did care, to some extent, about ending colonialism. He was a hypocrite because he did not care *enough*, not because he did not care *at all*. Hypocrisy does not consist simply in either self-regard or an absence of concern for one's professed values: it consists rather in an *excessive* self-regard and *insufficient* concern for one's professed values. This excess of self-regard leads hypocrites to fail, in many different ways, to properly honour the values they claim to have.

What are such failures like? Tartuffe represents the most extreme case, since *all* he seems to care about is deceiving others for profit. Still, there is a sense in which it is nevertheless false to say that Tartuffe did not care at all about piety. He certainly has no desire to actually *be* pious, but he does find the piety of others extremely valuable. He also values the appearance of piety in himself, since it allows him to advance his own interests. Tartuffe's problem is not that he does not value piety at all, but rather that he

values it for all the wrong reasons, and none of the right ones. He finds piety instrumentally valuable, and sees no intrinsic value in it, while the actual value of devotion to God is completely independent of whatever instrumental value it may happen to have. Tartuffe is a hypocrite because his desire to present a certain image of himself leads him to fail, in the most extreme possible way, to properly respect and engage with the values he claims to have.

Not all cases are as extreme as Tartuffe's, but other hypocrites display the same general pattern of failure. Consider, for example, Proust's Mme de Cambremer.[29] She fancies herself a sophisticated art critic with independent taste, and she confidently dismisses the Poussin in the Louvre. When she is told that Monet admires the same painting, however, she quickly changes her mind.[30] This change of mind reveals her hypocrisy: she cares more about *appearing* to be someone with sophisticated aesthetic taste than she actually cares about the paintings themselves.

It would be false, however, to say that Mme de Cambremer does not value art *at all:* indeed, she might care about it a great deal. After all, why would she value her reputation for aesthetic sensitivity if she did not think that aesthetic sensitivity was valuable itself? She really does care about art, but if she were less hypocritical, she would care much less about how her evaluations made her appear to others. Because she *does* care about her public image too much, she allows this to influence her evaluations, even though by her own lights the opinions of others should be irrelevant (after all, she values

---

[29] This example was originally raised in Bela Szabados and Eldon Soifer, 'Hypocrisy, Change of Mind, and Weakness of Will: How to Do Moral Philosophy with Examples', *Metaphilosophy*, 30 (1999), 60-78.

[30] Soifer and Szabados are particularly interested in what distinguishes hypocrisy from (mere) changes of mind.

her image as someone with independent taste). Like Tartuffe, Mme de Cambremer's desire to have a certain image leads her to fail to honour what she claims to value. Unlike Tartuffe, however, her claims to truly value art are not pure pretence.

IV. The self-image of hypocrites

Tartuffe and Mme. De Cambremer are radically different: one is a lying manipulator, while the other simply wants to be sophisticated but lacks the necessary confidence. Nevertheless, they share with many hypocrites a deep concern with their public image: what they care most about is how *others* see them. Many hypocrites aim to satisfy their desire for an image as someone with certain values by publicly asserting their allegiance to those values.

It might seem that hypocrisy requires this public aspect, and that it is essentially connected to a kind of blameworthy moralism, one aspect of which is an objectionable tendency to pass public judgments on the moral status of others when it is not one's place to do so.[31] Many hypocrites do indeed moralize in just this way, and one thing that makes hypocrisy objectionable is the public righteousness that often accompanies it. After all, one of the best ways to establish one's commitment to some value or other is to criticize the commitment of someone else to the same value.

Such public moralism certainly makes hypocrisy easier to diagnose: we can, after

[31] Robert Fullinwider argues that the vice of moralism is in adopting a position of authority that is not ours to take Robert Fullinwider, 'On Moralism', *Journal of Applied Philosophy*, 22 (2005), 105-20 at 111-2. And one of the most famous of all discussion of hypocrisy—St. Matthew's retelling of the Sermon on the Mount—opens with a warning against moralism: "Judge not, that ye be not judged," before continuing "thou hypocrite, first cast out the beam out of thine own eye; and then shalt thou see clearly to cast out the mote out of thy brother's eye." Matthew 7: 1-5

all, often easily spot breaks between the hypocrites public professions for having certain values and his actions with respect to those values. Nevertheless, such concern for one's *public* image is not essential to hypocrisy. It is possible for hypocrisy to be largely private: Orwell's humanitarian would still be a hypocrite if he kept his anti-colonial views to himself, and merely congratulated himself privately for having such an enlightened understanding of the world. Shyness is a barrier to moralism, but it is no obstacle to hypocrisy. So a hypocrite who is keen to publicly profess his values often displays *two* vices, adding moralism to his hypocritical concern with his own image.

That hypocrites need not be moralists reveals that not all hypocrites are primarily concerned with their *public* image. For many hypocrites, their *self*-image matters more than the image that *others* have of them. This influences the ways in which such hypocrites fail to honour their professed values. Orwell's humanitarian might care little about what others think of him, but still care very much about being able to see *himself* as morally righteous. This means that, unlike both Tartuffe and Mme de Cambremer, he will adopt moral beliefs that he truly believes are correct, and he will adopt them *because* he believes them to be correct. Both Tartuffe and Mme de Cambremer fail—albeit in very different ways—to honour the values they claim to have because they adopt those values for the wrong reasons: simply in order to deceive, in Tartuffe's case, and because someone she respects also has them, in Mme de Cambremer's. Orwell's hypocrite, however, may well have come to oppose colonialism for the most legitimate of reasons: because he saw that Britain's colonial subjects were unjustly suffering, and because he concluded that Britain had an obligation to alleviate the suffering it by abandoning the unjust practices of colonialism. These are perfectly legitimate reasons, apparently un-

driven by a self-interested concern with his image. What, then, is the failure of practical reason that makes Orwell's humanitarian a hypocrite?

He is a hypocrite because he does not properly honour the values he claims to hold. Though he can truly claim that he believes that colonialism is unjust and should be abolished, he has not fully integrated this belief into his larger pattern of beliefs, values, judgments, and intentions: his judgments and his actions do not properly reflect his beliefs in the way that they would if those beliefs were properly integrated. Were they integrated, he would feel shame at, as Orwell puts it, living a life of relative luxury by "robbing Asiatic coolies," and he would act in ways aimed at both rectifying injustice and minimizing his complicity in it. But he does not feel such shame, and he does not take such actions. He does not recognize the sacrifices he would have to make in order to bring about his stated political goal, and he may well be hesitant to make those sacrifices. He does not recognize the extent to which his own comfort depends in part on the ill-gotten gains of colonialism. His belief that colonialism ought to be abolished has not led him to see that this would require him to make sacrifices in his own life. He sees no problem with the luxuries he enjoys, and has no interest in giving them up. He is far too eager to criticize the moral complacency of others in failing to acknowledge the evil of colonialism, and not eager enough to consider the ways in which he, too, might be morally tainted. His reluctance to consider the ways in which he may be morally tainted can be attributed to his hypocritical self-regard: it is more important for him to *believe* that he is genuinely virtuous than it is to actually *be* virtuous. He believes that he is morally clean because he has the right opinions, and does not see the need to extend these opinions into concrete action.

The hypocrite who is concerned with his self-image displays a different sort of failure from those whose foremost concern is their public image. Tartuffe and Mme. De Cambremer acquire those values in the wrong way, and for the wrong reasons: their main concern is to adopt values that *others* see as correct. Orwell's hypocritical humanitarian, by contrast, seem to *form* his evaluative beliefs in the right way. Since he cares most about seeing *himself* as morally righteous, he holds the values he does because he thinks that they are correct, not because *others* do. His failure is that his judgments about how to act (and so, by extension, his actions) do not reflect those beliefs. So while the hypocritical humanitarian believes that colonialism is unjust, this belief is not integrated into his larger network of judgments, actions, and attitudes.

Such a hypocrite is in a puzzling state. Hume, discussing the beliefs of religious hypocrites, nicely characterized the problem: the purported belief is "some unaccountable operation of the mind between disbelief and conviction."[32] It seems unaccountable since, if he really does believe that colonialism is massively unjust, why do his actions not reflect this belief? Since they do not, it seems as if the hypocrite does not *really* believe what he claims to believe. (This is the thought that motivated the deception account.) This thought might be supported by an appeal to holism about belief, the view that "a belief is identified by its location in a pattern of beliefs."[33] According to the account of hypocrisy I favour, we can account for the hypocrite's puzzling state by noting that the belief that colonialism is wrong is *not* integrated into this larger network of beliefs (and other intentional states). That is what it means to say that a belief is "between disbelief

[32] David Hume, *The Natural History of Religion*, §12, Paragraph 15.
[33] Donald Davidson, 'Thought and Talk', *Inquiries into Truth and Interpretation* (Oxford: Clarendon Press, 1984), 155-70 at 168.

and conviction." According to holism, however, to say that a belief is not integrated into a network of other beliefs is to say that there is no such belief: we identify beliefs by their places in such networks. So it would appear that my argument runs foul of holism about belief, a position with powerful arguments in its favour.

This objection is too quick, however. It is true that the hypocrite's beliefs about colonialism are not fully integrated into his larger network of beliefs, but that does not mean it is not integrated at all. The hypocrite's belief that colonialism is wrong is integrated with his other beliefs about colonialism. If asked, he would agree that the suffering of the victims of colonialism is a bad thing, that their suffering is in part a result of the policies of the British government, and that a change in this policy would improve the lot of colonialism's victims. He thinks that these facts mean that the British government has a moral responsibility to do something—grant India its independence, stop exploitative trade policies, and so on. He thinks that those who fail to recognize these facts, or who do no believe that they create an obligation for the British government, are in error, and that their error represents a moral failing. He is even prepared to *act* on his beliefs, to some extent: he votes Labour, or perhaps even Communist, he reads 'enlightened' publications, and he tries to convert his friends and acquaintances to his way of seeing things. So there is a larger pattern of beliefs against which we can identify the hypocrite's views about colonialism. If this is right, then the hypocritical humanitarian *does* believe that colonialism should end, and he does so in a way that is consistent with holism about belief. The hypocrite's psychological state is not, in Hume's words, "unaccountable". Rather, it is between disbelief and conviction precisely because his belief does not lead him to draw the obvious conclusions about his

obligations, and so to *act* on those conclusions. Nevertheless, it is integrated enough to pass muster as belief.

The humanitarian's problem, then, is not that he does not believe that colonialism ought to be opposed, but rather that he does not fully integrate his beliefs into judgment and action. He has righteous disdain for the failure of others to both recognize and act on that obligation, and yet he does not extend the obligation to himself. All of his actions are, at best, symbolic—voting Labour, reading the right newspapers, and so on—and none of them make any material difference in either his own life or the lives of those he claims to believe he has an obligation to help. But this still leaves us with a puzzle: if he really does believe that he has an obligation, how to explain his failure to act on it?

V. Hypocrisy, akrasia, and failures of self-knowledge

There is nothing in principle mysterious about the idea of a rational agent failing to act as he believes he should. Akrasia is of philosophical interest in large part because it shows how such failures are possible. So perhaps akrasia can help to explain hypocrisy. It is quite possible, for example, that the hypocritical humanitarian's failure to give up the luxuries in his life that he ought to see are the unjust fruit of colonial exploitation was initially the product of akrasia.

Orwell's conviction that colonialism was unjust was largely the result of serving in the colonial police, and this conviction led him to resign his commission and, for a time, live the life of an ascetic. But we can imagine quite a different story: a member of the colonial police who becomes convinced of the injustice of colonialism, and so judges that he ought to resign his commission. When the time comes to do so, however, akrasia

gets the better of him, as he enjoys the luxuries that his position affords him. So, tempted by those luxuries, he acts contrary to his own considered judgments, and does not resign his commission. If his failure is genuinely akratic, then he will likely feel regret at his actions. Such regret could lead him to redouble his resolve and try again, but it could also lead in quite another direction. If he has a strong—perhaps hypocritical—desire to see himself as morally righteous, then this desire could have led him to interpret his actions in a way designed to reduce the feeling of shame and regret at following through on his values.

So perhaps, faced with a feeling of shame at his inability to resign, he comes to convince himself that, though colonialism is unjust, *he* is not a party to the injustice, since his role as a policeman is to *protect* colonial subjects from exploitation.[34] No doubt such rationalizations are common among those whose jobs directly confront them with injustice: it must be very hard to enforce laws one believes to be unjust. So the colonial policeman's akratic failure to follow through on his conviction that his job was unjust, combined with his excessive self-regard and a desire not to feel shame and regret at his actions, could have easily led him into self-deception about the justice of his own actions.[35] And, being self-deceived about the nature of his own actions, he can still confidently proclaim his belief that colonialism is an unjust institution. It is surely true that, in actual cases, hypocrisy, akrasia, and self-deception are often closely connected:

[34] Occupying armies do not claim that they will be greeted by cheering crowds merely as an exercise in domestic propaganda: it is likely that many of them actually believe it.

[35] It is important to note that all of this could have occurred unintentionally: the policeman's desire to see himself as morally righteous could have biased the ways in evaluated arguments about the harms of colonialism and his role in it without his *intending* to bias his evaluations. Mele outlines several of the mechanisms by which this can occur. Mele, *Self-Deception Unmasked* at Chapter 2.

our failings often reinforce one another, and are woven very fine.

Nevertheless, it is important to note that if the policeman's self-deception was successful, and he came to believe that there is nothing unjust with his role, then his continued employment would cease to be an example of akrasia, since it would no longer be contrary to his considered judgment. The more hypocritical he becomes, the less akratic his actions are. Akraisa might provide a causal explanation for hypocrisy, but genuine hypocrites are not akratic.

This imagined transition from akrasia to hypocrisy highlights the difference between the two: akratics know, in some sense, that they have acted contrary to their better judgments, and therefore experience regret. If this regret leads to self-deception, however, their regret disappears, because they no longer know that they acted contrary to their own judgments. The difference, in other words, is one of self-knowledge. Most hypocrites, unlike akratics, do not know that they are hypocrites.[36] But how is their lack of self-knowledge to be explained?

One possible answer to this question is that hypocrites lack self-knowledge because they are self-deceived about the content of their attitudes; they do not know what they believe. This account of hypocrisy does not work, not simply because it mischaracterizes *hypocrites*, but because, as I will argue in Chapter 5, self-deception about the content of one's own attitudes is not possible for *anyone*. Nevertheless, garden-variety empirical self-deception can both emerge from and lead to hypocrisy, and can contribute, in important ways, to the hypocrite's lack of self-knowledge.

[36] A different way of making this point: akrasia has a characteristic phenomenology, but hypocrisy does not.

Hypocrites display an excessive self-regard: they care more about their image for having certain values than they do about the values themselves. This self-regard can lead directly to adopting the sorts of motivationally biased cognitive strategies that generate non-intentional self-deception.[37] Compare two people who profess to value the environment, one hypocritically and one sincerely. The difference between them is not whether they value the environment: both of them want to be eco-friendly. The difference is rather how they rank the importance of the environment compared to other related values. The hypocrite cares more about *believing* that she is eco-friendly than she does about actually *being* eco-friendly, while the sincere environmentalist's preference are reversed. This will affect the ways in which they examine the evidence in a way that can lead to self-deception.

In gathering and interpreting evidence, we not only aim to form true beliefs; we also aim to minimize the chance that we will form false beliefs that are particularly costly. The errors that we find most costly can reveal what we value most. Both the hypocrite and the sincere environmentalist want to believe that they are eco-friendly, and both want this belief to be true. They differ, however, in which false beliefs they are most concerned with avoiding. Since the hypocrite cares most about his image, the error she would find the most costly would be to falsely believe that she was not eco-friendly. The sincere environmentalist, on the other hand, would prefer to actually *be* eco-friendly and falsely believe the opposite. The hypocrite most wants to avoid a bad self-image, and the sincere environmentalist most wants to avoid harming the environment. These preferences can influence the ways in which they gather and interpret evidence.

[37] I will discuss these strategies in greater detail in Chapter 5.

It is easy to see how these preferences can lead the hypocrite into self-deception. Imagine that she takes pride in her use of bio-fuels in her car, believing that they are a carbon-neutral alternative to fossil fuels. She is them confronted with a report arguing that, in addition to diverting important crops from the food-chain by feeding corn to cars rather than to starving people, over their lifespan bio-fuels in fact release *more* greenhouse gasses into the atmosphere than fossil fuels. Her preferences will influence the way in which she treats this information. Since she wants, above all, to avoid falsely believing that she is harming the environment, she is more likely sceptical of the report, look for reasons to doubt its authenticity, assume that it is the work of the fossil fuel industry, or that it is propaganda from global warming denialists. Her preferences will influence the sorts of evidence that she is willing to consider, the hypotheses that strike her as likely (and those that simply fail to strike her at all), and so on. In general, her motivational bias will lead her to employ cognitive strategies that make it unlikely she will come to believe the claims of the report, since her threshold for believing that the report is true will be much higher than her threshold for believing that it is false.[38] If, in fact, the claims made in the report are true, and a less biased person would have seen that there were good reasons for believing those claims, then her bias will have led her to form a false belief in the face of the evidence, and so she will be self-deceived.[39]

---

[38] This is an illustration of Mele's claim self-deception can arise in part because people "test hypothesis in ways that seem natural to them in the circumstances, and a major part of what makes their tests seem natural is relevant desire-shaped error costs." Mele, *Self-Deception Unmasked* at 42. Mele argues that need be nothing intentional about the influence of desire on hypothesis testing.

[39] The sincere environmentalist might also become self-deceived, if his love of the environment leads him to trust the report when he has good reason to doubt it. But this self-deception would not be hypocritical, since it would have been brought about by, in

The hypocritical environmentalist's self-deception about bio-fuels contributes to her hypocritical failure to properly honour the values she claims to hold. By being self-deceived, she will continue in a practice—using biofuels—which she ought to know is inconsistent with her purported environmentalism. Moreover, her self-deception emerged from her excessive concern for her own image, and since it leads her to fail to properly honour her values, her self-deception is a central part of her hypocrisy.

In Chapter 5, I will argue that the nature of first-person rational agency means that it is not generally possible to be deceived about one's own attitudes. There is an asymmetry between the first and third person perspective on our mental lives, and one aspect of the asymmetry is that, when it comes to the content of our own attitudes, we have immediate, non-evidential access to what others can access only by observation. There is, however, another side to the self/other asymmetry. There are facts about us—including facts about our attitudes—that *others* are in a better position to know that we are, and there are even facts about us and our mental states that *only* others are in a position to know. For example, I cannot know which of the beliefs that I currently hold are false, even though other people can often easily identify my false beliefs.[40] Other people can know that one of my beliefs is false, even if *I* cannot. In a similar vein, Socrates argued that I cannot knowingly choose the bad, although others can know that I

---

effect, a desire to avoid hypocrisy. The self-deceived but sincere environmentalist cares about the environment so much that he takes 'better safe than sorry' to unreasonable lengths, while the hypocrite cares about herself so much that she unreasonably rejects the 'better safe than sorry' approach.

[40] This yields Moore's paradox: I cannot truthfully say "I believe that it is raining, but it is not raining", even though it might be *true*, and other can say it of me.

do.[41]

Hypocrisy appears to be another case where there are facts that are accessible from the third person perspective but which are inaccessible from the first-person perspective. One of the frustrating aspects of hypocrites is that the very behaviour that others recognize as evidence of hypocrisy—say, the environmentalist's self-deceived insistence that bio-fuels are environmentally friendly—she falsely sees as evidence of her virtue. This suggests that the hypocrite's self-deception about empirical matters—such as the environmental effects of bio-fuels—can contribute to a deeper level of self-deception *about herself*.

Others are almost always in a better position to judge whether someone is hypocritical than the hypocrite himself: we know that the environmentalist is a hypocrite while she believes she is virtuous; we know that Orwell's humanitarian's commitment to his cause is weak even if he does not. With the exception of clear-eyed hypocrites like Tartuffe, most hypocrites will genuinely not believe that they are hypocritical. This contributes to the all-too-common phenomenon of hypocrites righteously accusing one another of hypocrisy, unaware that they are displaying the same failure for which they chastise others.[42] In fact, I will argue below that many hypocrites are deeply deceived

---

[41] Plato, *Protagoras*. For an argument that we *can* knowingly choose the bad, see, for example, Michael Stocker, 'Desiring the Bad: An Essay in Moral Psychology', *The Journal of Philosophy*, 76 (1979), 738-53.

[42] This form of "meta-hypocrisy" (hypocritical accusations of hypocrisy) is especially common in political debate. A recent example: by way of responding to an EU threat to ban the import of seal products as a way of protesting the seal hunt, a Newfoundland MP suggested that Canada should ban the import of the products of the German wild boar hunt. While the MP was surely right that there is something hypocritical about Europeans protesting the hunt of cute mammals in North America while allowing the hunt of ugly mammals in Europe, his suggested counter-ban was equally hypocritical. After all, if he

about themselves, and even about their own attitudes, while maintaining that no one can be self-deceived about the *content* of her attitudes.

The hypocrite's failure of self-knowledge comes from falsely ascribing virtues to herself—she believes she has virtues she does not have. It may be that some such self-ascriptions are self-falsifying. As Williams points out, "the modest person does not act under the title of modesty"; no one who boasts of being modest can possibly *be* modest.[43] Not all self-ascriptions of virtue, however, are self-falsifying in this way. It is possible to properly describe oneself as just without betraying a misdirection of ethical concern. Even if they are not self-falsifying, however, self-ascriptions of virtue can certainly be *false*. Those who confidently ascribe virtues to themselves suggest to us that their moral attention is misdirected, since the virtue terms do not generally occur in the deliberations of genuinely virtuous agents. They suggest, in other words, that it is his concern for his image, rather than his purported values, which moves him to act. To confidently assert that one is generous, or courageous, for example, seems to display the kind of misdirection of concern that is inconsistent with a genuine possession of virtue. Such false self-ascriptions of virtue lie at the heart of the hypocrite's failure of self-knowledge.

False self-ascriptions of virtue can happen in two ways. First, someone can falsely believe that his actions meet the standard of the virtue that he claims to have. Second, he can falsely believe that his attitudes are virtuous. The first sort of mistake is relatively straightforward. The standards for what counts as an instance of any particular virtue are independent of an agent's beliefs about what those standards are, and so my belief that

---

thinks that the seal hunt is justifiable, then he should have no problem with the boar hunt either.

[43] Williams, *Ethics and the Limits of Philosophy* at 10.

my actions are generous, or brave, does not guarantee that I am right. Such false beliefs can easily be the result of self-deception, since a hypocrite can form false beliefs about the standard of virtue in much the same way that the environmentalist was self-deceived about the facts about bio-fuels.

The second sort of mistake—falsely believing one's attitudes are virtuous—is more complicated. After all, it involves having false beliefs about one's attitudes, and I argue that we cannot be self-deceived about the content of our own attitudes. The mistake at issue here, however, is not about the *content* of the hypocrite's attitudes, but about the proper *evaluation* of those attitudes. Just as hypocrites can be wrong about whether their *actions* meet the standard of virtue, they can also be wrong about whether their *attitudes* are part of a virtuous or praiseworthy deliberative outlook. If I do not truly understand the virtue of generosity, then I may believe that a self-congratulatory attitude towards my apparently generous acts is a praiseworthy sign of generosity, rather than a failure to be truly generous. Virtuous agents take as reasons considerations that non-virtuous agents fail to recognize, and so virtue involves, not simply being motivated by certain considerations, but a deliberative outlook that involves a sensitivity to their status as reasons that non-virtuous people lack. Virtue, in other words, involves a distinct domain of knowledge.[44] It is therefore no surprise that those who are not virtuous can have false beliefs about the sorts of reasons for action and attitudes that are consistent with virtue.

---

[44] This is a central claim in McDowell's argument in both 'Virtue and Reason' and, more indirectly, in 'Non-cognitivism and rule-following.' For two other compelling arguments that the "should" of practical reason just is the "should" of ethical virtue, see Philippa Foot, 'Rationality and Virtue', *Moral Dilemmas* (Oxford: Clarendon Press, 2002), 159-74, and Kieran Setiya, *Reasons without Rationalism* (Princeton: Princeton University Press, 2007).

Both mistakes—about actions and attitudes—are, for the hypocrite, products of excessive self-concern. Their excessive focus on their image makes them too concerned that their actions and attitudes be understood as virtuous, and not concerned enough with how they ought to act. In primarily directing their attention to whether or not they are in fact virtuous, they focus on the wrong sorts of considerations, considerations that have no place in the deliberations of the truly virtuous. The hypocrite's excessive self-concern therefore leads directly to false beliefs about his own virtue, and so to an important failure of self-knowledge.

VI. Praise and blame for hypocrites

On the deception account, hypocrites stand "twice condemned," once for having objectionable moral views that they dare not reveal, and a second time for deceiving others about those views.[45] In fact, some proponents of the narrow view find it obvious that the hypocrite even *more* despicable than avowed egoists.[46]

---

[45] Feder Kittay thinks that the puzzle about hypocrisy is in explaining "why the sexist … who is hypocritical is usually twice condemned" while the avowed sexist is only condemned once. Feder Kittay, 'On Hypocrisy', at 277.

[46] Szabados and Soifer claim that, when we compare the avowed egoist to the hypocrite, "intuition strongly suggests that the hypocrite is the worse offender, and the long history of condemnation of hypocrisy indicates that this is a common perception." Given that egoism has an equally long history of condemnation, their intuition seems poorly supported, but the idea is that hypocrites are condemned once for having bad values, and a second time for deceiving others about them. Szabados and Soifer, *Hypocrisy: Ethical Investigations* at 172. And MacKinnon argues that the hypocrite is a deceiver with a desire to undermine morality for her own advantage, and is blameworthy on both counts. Mackinnon, 'Hypocrisy, with a Note in Integrity.' Both Szabados and Soifer and MacKinnon argue that our double condemnation of the hypocrite is a serious problem for consequentialism, since we blame hypocrites even when their actions have good consequences. For a reply on behalf of consequentialism, see William Shaw, 'Is Hypocrisy a Problem for Consequentialism?' *Utilitas*, 11 (1999), 340-46. There is an

Even those who reject central elements of the narrow account of hypocrisy seem to accept this account of the blameworthiness of hypocrisy.[47] They agree with the defenders of the narrow view that the wrongness of hypocrisy is closely connected to deliberate deception—they just disagrees that hypocrisy is always characterized by deliberate deception, and this leads them to conclude that hypocrisy is not always wrong.

But this ignores another important possibility: hypocrisy is not essentially characterized by deception, *but it is blameworthy nonetheless*. Some hypocrites, like Tartuffe, are blameworthy for the very reasons that the deception account identifies. Most hypocrites, however, are not. Their failure is that they misdirect their ethical attention: they care more about their image for having certain values than they do about the values themselves. This misdirected ethical attention is not the same as a complete lack of concern for morality, let alone adopting the goal undermining it. Hypocrites do not all

---

analogy here with Benett's comparison of Huck Finn to Himmler: for Bennett, Huck is in some sense *worse* than Himmler, since he has similar deplorable moral beliefs *and* he is weak and irresolute. Jonathan Bennett, 'The Conscience of Huckleberry Finn', at 45.

[47] Daniel Statman, for example, argues that, since hypocrisy is often characterized by self-deception, and since self-deception is involuntary, we should be less quick to blame hypocrites Statman actually equivocates about whether self-deception is intentional. He portrays self-deception as a strategy for carrying out the interpersonal deception that narrow hypocrisy requires, but he also counts it as involuntary, and so as deserving pity rather than blame. It is hard to see how these two positions can co-exist, since self-deception can only be a *strategy* if it is intentional. Similarly, Dan Turner rejects the association of hypocrisy with deception. He argues instead that hypocrisy is simply characterized by a disparity or conflict in values. We express values in our beliefs, our words, and our actions, and so hypocrisy requires a "disparity pair": a belief in conflict with an action, or a pretended belief in conflict with a genuine belief. That is all hypocrisy requires—there is no reference at all to deception, let alone wilful deception. Turner concludes that, since hypocrisy does not require deception, "the evil of hypocrisy is on par with the evil of logical inconsistency," which is to say, not that evil at all. Dan Turner, 'Hypocrisy', *Metaphilosophy*, 21 (1990), 262-69 at 266. In addition to being unable to explain the wrongness of hypocrisy, Turner's account seems unable to distinguish hypocrisy from other forms of moral inconsistency such as akrasia and moral cowardice.

lack moral seriousness—many take morality very seriously indeed, but have a deeply mistaken view about what is ethically important. The hypocrite's failure is not always a lack of sincerity, but an excessive concern with their own moral assessment, and so a failure to properly direct their concern.[48] This failure can take many forms: hypocrites can be complacent, failing to even consider the demands of the values they claim to have; they can be self-deceived, too quick to adopt beliefs that allow them to maintain a positive self-image; and they can overly sophistical and clever, rationalizing almost any behaviour as consistent with their values and finding almost any excuse to doubt the validity of purported ethical obligations. Each form of misdirection is blameworthy, since each represents a failure to properly engage with morality. Hypocrites can fail to put in the often hard work of moral reflection, they can fail to face up to the real cost of their moral commitments, and so they can fail to fully integrate their moral values and beliefs into a larger network of beliefs, intentions, and desires. These failures can be due to a lack of willingness to try, but it need not be. It is possible to try to get things right and still fail: moral reflection can be hard, and failures of self-knowledge and misdirected ethical attention can trip up even those whose commitment to morality is sincere.

Hypocrites therefore display real moral failures, and they are certainly blameworthy. But, though hypocrisy is blameworthy, it is a mistake to see it as the only unforgivable sin, "twice condemned", and even worse than vicious egoism. A hypocrite

[48] This is where Roger Crisp and Christopher Cowton's otherwise admirably nuanced account of hypocrisy goes wrong. They argue that hypocrisy is characterized by a lack of moral seriousness, and so they fail to see how hypocrites can be quite *serious* about morality, but wrong about what is morally significant. Roger Crisp and Christopher Cowton, 'Hypocrisy and Moral Seriousness', *American Philosophical Quarterly*, 31 (1994), 343-8.

who acts well in one context but not in another need not be condemned in both. The environmentalist who persists in the use of bio-fuels is certainly blameworthy for her hypocritical self-deception: she ought to know better, and so to change her behaviour. But if she is also dedicated to composting, recycling, eating locally, and donating money to worthy environmental causes, she can also be praised for her genuine environmentalism. These actions can be undertaken for all the right reasons and with an appropriately praiseworthy deliberative outlook. They are not empty gestures, and her environmentalism is not a mere pretence to virtue, deserving of censure. She might genuinely care about the environment; her hypocrisy is in sometimes not caring about it nearly *enough*, despite her valuing image for caring about it a great deal.

Instead of thinking of hypocrites as malicious liars, out to subvert morality, it is better to think of them as closely analogous to akratics. Both are, in Aristotle's terms, "correctible." They are blameworthy (though not always: see Chapter 3), but not vicious. Hypocrites are unlike akratics in that they do not know that they are hypocritical, but this does not mean that all hypocrites will resist the accusation of hypocrisy. Many will of course rationalize away the accusation, but others may treat it as a genuine discovery and be moved to self-improvement. Hypocrites, like the rest of us, have attitudes that are under their control: they can bring those attitudes into alignment once they gain the proper self-understanding. Bringing their hypocrisy to their attention can be an important step in developing this self-knowledge. Our reaction to hypocrites who fail to live up to the standards they proclaim should be "your principles are the right ones, but you can do much more to honour them," not "you're an unprincipled sham." The fact that they can and do have the right values, even if they do not fully integrate those values into their

network of attitudes, means that hypocrites can be improved, can be made to see the light. Like Huck Finn, the hypocrite's moral blindness is often only partial. Both Huck and the hypocrite can be praised for the moral vision they do have, even as they are criticized for their failure to fully extend that vision.

I have criticized the deception account for failing to properly capture hypocrisy's scope. It might be objected that my account similarly fails to account for the extreme condemnation that hypocrisy received. Such an objection would be misplaced, however. On my account, hypocrisy can be a serious vice. I simply maintain that many vices are worse, that hypocrites do not all aim to undermine morality, and that hypocrites can merit praise for actions that express the very values about which they are hypocritical. Moreover, we should take many accusations of hypocrisy with a healthy dose of salt. They are often forceful not because hypocrisy really does merit the strongest possible condemnation, but because many accusations are themselves exercises in hypocrisy: one of the best ways for a hypocrite to demonstrate the strength of his purported commitment to morality is to question the sincerity of someone else's. The prevalence of widespread condemnations of hypocrisy does not mean that hypocrisy is as blameworthy as it role in the cultural conversation might lead us to believe; rather, it means that our cultural conversation is often highly hypocritical.

Chapter 5: When Self-Knowledge Fails

Emma likes to think of herself as something of a matchmaker. She has recently dedicated herself to getting Mr. Elton to love her friend Harriet, and is convinced that she has succeeded. When Mr. Knightly, Mr. Elton's close friend, tells her that the latter would never marry someone of Harriet's background, Emma is initially quite vexed. But in the end, she does not believe him. Allow Jane Austen to take up the story:

> She was not so materially cast down, however, but that a little time and the return of Harriet were very adequate restoratives… He had frightened her a little about Mr. Elton; but when she considered that Mr. Knightly could not have considered him like she had done, neither with the interest nor (she must be allowed to tell herself, in spite of Mr. Knightly's pretensions) with the skill of such an observer on such a question as herself, that he had spoken it hastily and in anger, she was able to believe, that he had rather said what he rather wished resentfully to be true, than what he knew anything about.[1]

Austen's delightful description makes it quite clear that Emma is hopelessly self-deceived. Her self-image as a skilled matchmaker leads her to overrate her own insight into Mr. Elton's emotional state, to count his clear displays of interest in *her* as evidence of his love for Harriet, and to miss the abundant evidence of his lack of affection for Harriet. Emma is "too eager and busy in her own previous conceptions and views to hear [Mr. Knightly] impartially."[2] She is so unaware of her self-deception that she even accuses the clear-eyed Mr. Knightly of being self-deceived.

Emma—like Huck Finn and the hypocritical environmentalist and humanitarian—lacks self-knowledge. Emma thinks that she is a skilled matchmaker and an insightful

[1] Jane Austen, *Emma* (New York: Bantam Classic, 1981) at 61-2.
[2] Ibid. at 102.

observer of human nature; Huck believes that he is a bad boy, and destined to remain that way; Orwell's humanitarian believes that he virtuously lives up to the values he claims to have. In all three cases, we know better. We can see that Emma is shallow and self-obsessed, that Huck is a good boy, despite his akrasia, and that the hypocrite is not nearly as virtuous as he believes. Such examples illustrate a general point: other people often know us much better than we know ourselves. But how is this possible? How could someone else, whose knowledge of me can only ever be indirect and incomplete, possibly know me better than I know myself?

For Emma, Huck, and the hypocrite, this is a question of the first importance. For each of them, moral development depends in large part on coming to see that (and how) their own self-assessments have gone wrong. This means that each of them will need to acquire self-knowledge that they currently lack. To see how this is possible, we need to understand exactly how it is that they have gone wrong: what is the source of their lack of self-knowledge? And what, exactly, are Emma, Huck, and the hypocrite mistaken about in being mistaken about themselves?

One obvious suggestion is that failures of self-knowledge are the product of self-deception. We all have a strong interest in our own self-image, and this interest can have a distorting effect on the way we see ourselves. Recent discussions of self-deception have emphasized the connection between these two conditions by arguing that self-deception *requires* a lack of self-knowledge. As Richard Holton puts it, self-deception is "more

concerned with the self's deception *about* the self, than with the self's deception *by* the self."[3]

That self-deception can involves a lack of self-knowledge is an important and easily overlooked point. At a minimum, the self-deceived do not know that they are deceived. But how should we understand this lack of self-knowledge? What precisely is it that we are wrong about, when we do not know the truth about ourselves? Emma certainly does not know that she is self-deceived, but perhaps her lack of self-knowledge goes much deeper. I have described Emma as believing that Mr. Elton loves Harriet, but this description may be tendentious; maybe Emma's self-deception means that she only *thinks* that she believes this, while believing 'deep down' that he has no interest in Harriet. Emma claims to believe that Mr. Elton loves Harriet; the hypocrite claims to believe that colonialism is unjust and ought to be abolished; Huck claims to believe that he has an obligation to turn Jim in. But perhaps they are all mistaken about what they believe—perhaps the right thing to say to them is "you might *think* you believe that, but you're wrong." That is, perhaps their real error is in having false second-order beliefs—beliefs about what it is that they believe—and not first-order false beliefs, since they do not in fact have any such beliefs, despite their second-order beliefs to the contrary. This suggestion is reflected in the folk-psychological language we use to describe others: without being in the grips of a theory, we often say that someone "doesn't know what he believes", or that a misguided friend "only thinks that's what she wants." David Patten has recently advanced an argument to this effect, which claims that those who are self-

[3] Richard Holton, 'What Is the Role of the Self in Self-Deception?' *Proceedings of the Aristotelian Society*, 101 (2001), 53-69 at 53. Emphasis in original.

deceived are often deceived about the *content* of their own minds.[4] If Patten is right, then self-deceived agents can be wrong about what it is that they believe.[5]

In this chapter, I argue that such failures of self-knowledge are not possible. Those who are self-deceived fail, in some important respects, to know their own minds, but they cannot be self-deceived about the content of their own attitudes. Their failure is not in knowing what they believe, but in understanding themselves. The most profound failures of self-knowledge do not concern the content of our attitudes, but the quality of our characters. This has important effects for our understanding of moral education and in just what is involved in the aspirations of morally flawed agents to a greater level of virtue.

I. Accuracy of belief v. content of belief

There are certainly many facts about me that others can be in a better position to know than I am. If am hurt but do not know if I have a broken nose and a concussion, I can go to the doctor, and she can tell me. Such failures of self-knowledge can result from self-deception: one of Donald Davidson's most-discussed examples is a bald man who is self-deceived about his own baldness.[6] But Patten's claim is much more radical. Standard cases of self-deception involve a failure to recognize that one's beliefs do not accurately

[4] David Patten, 'How Do We Deceive Ourselves?' *Philosophical Psychology,* 16 (2003), 229-46.

[5] In the text, I will generally speak of beliefs and set aside the complexity of the other cases. The argument is quite general, however, and applies, not only to belief, but to all of what T.M. Scanlon calls the "judgment sensitive attitudes." These are attitudes "for which reasons in the standard normative sense can be offered" and include not only beliefs, but also desires, intentions, and many emotions. T.M. Scanlon, *What We Owe to Each Other* at 20-24.

[6] Donald Davidson, 'Deception and Division' at 199.

reflect the facts. Patten suggests that in many cases the mistake goes much deeper, and that the self-deceived fail to even know what it is that they believe. That is, he argues that I can be self-deceived not only about whether my beliefs are *true*, but also about what my beliefs *are*. This is a mistake of a completely different order, not simply about the *accuracy* of my beliefs, but about their actual *content*. In other words, it is a second-order false belief about what it is I believe, rather than a first order false belief about some empirical fact. If Patten is right, then the possibility of a failure of self-knowledge extends very deep indeed, and there are simply no facts about us about which we cannot be mistaken, or that others might not be better placed than us to know.

Patten claims that the difference "between what the [self-deceived] individual believes, or wants, and what he believes himself to believe, or want"[7] can explain "the psychic tension that is often associated with being in a state of self-deception."[8] He argues that we can have false beliefs about our own attitudes because we sometimes form beliefs about our motives in acting, and so about our beliefs and desires, in the same way that we form those beliefs about others: we infer them from the evidence of our behaviour. When these inferences go awry, we can be the victims of self-deception. But are such radical failures of self-knowledge possible?

If we can be self-deceived about what we believe, then it should be possible for someone to sincerely assert that she believes that p and to be mistaken about the truth of that avowal. She thinks that she believes that p, but she does not. There are two problems with this view. I take up the second, whether such self-deception is even possible, in

[7] Patten, 'How Do We Deceive Ourselves?' at 230.
[8] Ibid.

Section IV, below. The first problem is epistemological: even if it were possible, how could we ever know it was true?

Imagine someone asserting that she believes eating meat is cruel. It is tempting to suppose that the evidence of her behaviour might provide us with clues as to whether or not she really believes this: if, despite her assertion, she eats meat, or works in a slaughterhouse, we might conclude that, no matter what she says, she does not really believe that eating meat is cruel.[9] This is not enough for us to conclude that she is wrong about what she believes, however. After all, she could simply be lying. So what we need is a method for telling *both* that she is not lying *and* that she does really not believe that eating meat is cruel—a way of knowing that she is sincere, but mistaken. If there is such a method, then we may have a way of knowing that someone is mistaken about what she thinks she believes.

The problem is that no such method exists. Any evidence that would suggest that she is not lying—say, a refusal to eat meat, or membership in PETA—is at the same time evidence that she really does believe meat-eating to be cruel. And any evidence that she is lying—meat-eating, recreational hunting—is also evidence that she is mistaken about the content of her beliefs. Short of a confession, anything that would suggest that she is lying would *also* suggest that she is insincere, and any evidence that suggests that she is sincere also suggests that she is not lying.[10] There comes a point, says Anscombe, "at

[9] In fact, even this is not quite true: perhaps she believes it is cruel, but she *values* cruelty.
[10] This is a problem with the reliability of testimony: inferences from testimony is unreliable not simply because all inferences from evidence is unreliable, but also because assertions are free, intentional actions aimed at inducing beliefs, and those who make assertions know this. So we need to guard against the possibility that they are *intending* to mislead us, in addition to the possibility that they are simply mistaken. When the

which the skill of psychological detectives has no criteria for its own success."[11] No matter how many clues the psychological detective uncovers, she can never know what they show—in other words, they are not genuine clues. In fact, even a confession would not resolve the matter. Confessions can be false, and so the problem is simply pushed back a level. There seems to be no evidence that would leave us confident that her assertion about the content of her beliefs was both sincere and mistaken: our confidence that it is mistaken must diminish as our confidence that it is sincere increases.

I propose to set this problem aside. As Patten admits, we might be *incorrigible* with respect to our attitudes even if we are not *infallible*: perhaps no one is ever in a better position to judge their contents than we are ourselves, but this does not mean we cannot be mistaken.[12] The problem of knowing whether someone is mistaken about her own beliefs may just be a particularly acute instance of the problem of other minds, which is an epistemological problem. Those who find sceptical arguments about other minds convincing need not believe that other minds do not exist: their scepticism is about the warrant for that belief. Moreover, Patten does not just suggest that we can be *mistaken* about what we believe: he suggests that we can be *self-deceived*, and this is a much stronger claim. I will argue that, as it is standardly understood, such self-deception is impossible—though self-deception can lead to deep failures of self-knowledge, we

---

assertion is about what the speaker *believes*, however, we cannot disentangle these two possibilities. For a discussion of some of the problems of the role sincerity plays in forming beliefs on the basis of testimony, see Richard Moran, 'Problems of Sincerity', *Proceedings of the Aristotelian Society,* 105 (2005), 341-61.

[11] G.E.M. Anscombe, *Intention* at § 27.

[12] Patten, 'How Do We Deceive Ourselves?' at 230. Moran makes a similar point: "sincerity does not function as a guarantee of access to the speaker's beliefs." Moran, 'Problems of Sincerity', at 358.

cannot be self-deceived about what we believe. Our failures to know ourselves must be explained in some other way.

II. Is self-deception intentional?

A self-deceived agent is one who believes something false despite having good reason for believing that it is true. The self-deceived agent's mistake is standardly attributed to the presence of a desire: she believes something that is false because she *wants* it to be true. But a desire that p be true is no reason at all for *believing* that p is true. How, then, can self-deception work? The two dominant accounts of self-deception disagree over whether the self-deceived agent's desire leads her to *intentionally* adopt the false belief that p is true. Patten breaks with both camps in rejecting the view that desire is necessary for self-deception. As we shall see, all three nevertheless agree about the mechanisms by which self-deception is achieved.

For intentionalists about self-deception, such as Donald Davidson and David Pears, an agent is self-deceived if he is motivated to believe that ~p, believes that p is more likely than ~p, and so acts "with the *intention* of producing a belief in the negation of p."[13] Self-deception is more than an innocent mistake: it is, on this view, an intentional mistake, and as such a clear failure of rationality, since it is a belief adopted despite, and indeed *because* of, clear evidence that it is false.

This obvious irrationality presents intentionalists with something of a paradox: their view requires that the self-deceived agent believes that p and *therefore* believes that ~p. Davidson's example is of a bald man who believes that he is not bald *precisely*

[13] Davidson, 'Deception and Division', at 208. Emphasis added.

because he believes that he is bald, and he finds this belief distressing. But this is quite odd, since the true belief seems to *sustain* the false one. This implies belief in a contradiction, and also seems to require that I could acquire a false belief at will. But as Bernard Williams points out, "If in full consciousness I could acquire a 'belief' irrespective of its truth, it is unclear that… I could I could seriously think of it as a belief."[14] Intentional self-deception, then, seems to violate the very logic of belief.

The solution to this paradox that intentionalists adopt is to subdivide the self-deceived mind, sometimes by positing the existence of intentional sub-agents.[15] This homuncular strategy dissolves the paradox of intentional self-deception by making self-deception analogous to interpersonal deception, which is uncontroversially intentional. But the homuncular solution is problematic: why, for example, does the deceived main system accept the beliefs that are generated by the deceiving sub-system, of which it is unaware? As Mark Johnston points out, the answer cannot be that it *colludes* with the deceiving sub-agent, since that raise the same intentional paradoxes that the homuncular strategy was intended to explain away.[16] The ways in which a divided mind can carry out the complex process of self-deception is left unexplained.

The deflationist account of self-deception, most notably defended by Mark Johnston and Alfred Mele, reject this intentionalist picture of self-deception. They point

---

[14] Bernard Williams, 'Deciding to Believe', *Problems of the Self* (Cambridge: Cambridge University Press, 1973), 136-51 at 148.

[15] Davidson claims that there are "boundaries between parts of the mind" of the self-deceived person. Davidson, 'Deception and Division' at 211. Pears argues that there is a separate centre of agency within the self-deceived mind. David Pears, *Motivated Irrationality* (Oxford: Oxford University Press, 1984).

[16] Mark Johnston, 'Self-Deception and the Nature of Mind', in Brian Mclaughlin and A.O. Rorty (eds.), *Perspectives on Self-Deception* (Berkeley: University of California Press, 1988), 63-91 at 84.

out that intentionalists are only driven to propose problematic homuncular accounts of self-deception because they cling to the paradoxical claim that self-deception must be intentional. Once the insistence that self-deception is intentional is abandoned, however, the paradoxes dissolve, and self-deception becomes relatively easy to understand.[17]

As Mele points out, when one's desire that something be true leads one to falsely believe that it is true, it need not do so "as part of an *attempt* to deceive oneself, or to cause oneself to believe something, or to make it easier for oneself to believe something."[18] Desires are standardly understood as explaining actions by providing us with *reasons* to act, but they can play other roles as well. Mele claims that desires can also cause us to misinterpret evidence, to selectively focus our attention, or to selectively gather evidence, all without it being true that we do so as part of an attempt to acquire a false belief. Similarly, for Johnston motivated beliefs are the non-intentional outcomes of a "mental tropism," which is a "nonaccidental mental regularity" that is not reason-governed.[19] The mental process responsible for self-deception aims at the reduction of anxiety rather than at true beliefs.

[17] It is worth noting that deflationists accept that it is possible to intentionally set out to acquire a false belief. I may, for example, take a drug that induces retroactive amnesia (as some sleeping pills do), and arrange things so that, when I wake up, I am misled about my activities prior to taking the pill. There is nothing paradoxical about such a case, but in being consciously calculated, it is decidedly rare and non-standard. Mark Johnston argues that such cases involve the use of "autonomous means" to achieve the desired belief, and so do not count as genuine self-deception. Ibid. at 76-7.

[18] Alfred Mele, *Self-Deception Unmasked* at 18.

[19] Johnston, 'Self-Deception and the Nature of Mind', at 66. A tropism is an example of the non-voluntary mental causation that Anscombe identified.

Even intentionalists like Davidson allow that, in cases of causal deviance, a desire to do something can cause me to do it without my intending to do it.[20] In fact, a desire to act in one way can even cause me to act in exactly the *opposite* way, as when a strong desire to act calm, cool, and collected makes it all the more certain that I will blush, stammer, and drop my beer. So it is generally agreed that desires can play a *causal* role in action without playing a *rationalizing* role. Deflationists argue that desires can play the same role in explaining belief-formation as they do in explaining action. Desires can have an effect on what I believe without my ever taking them as a *reason* for adopting one belief rather than another. Desires can lead me to mishandle evidence, and therefore adopt false beliefs, all without it being true that I was *trying* to adopt a belief I knew to be false. In these cases, desires explain why we are self-deceived, but they do not give a reason (in the sense of a rational explanation) for self-deception.[21] If this is right, then self-deception need not be intentional.

The intentionalists and the deflationists disagree about whether self-deception can be intentional, but they agree that it is set in motion by a desire. Patten disagrees with both camps, rejecting the idea that a desire is necessary for self-deception. His argument relies on the claim that we are sometimes unsure about our own motives. When we are,

---

[20] Donald Davidson, 'Freedom to Act', *Essays on Actions and Events*.

[21] Of course, for Davidson reasons *are* causes of a certain sort, so rationalizing explanations are a species of causal explanation. See Donald Davidson, 'Actions, Reasons, and Causes'. Mele's point is that, in self-deception, desires play a causal role *without* playing a rationalizing role. As Anscombe points out, "mental causality is not restricted to choices or voluntary or intentional actions, but is or wider application… it includes some involuntary actions." Anscombe, *Intention* at § 12. Another powerful argument that desires can cause actions without the intermediary of an intention can be found in the work of David Velleman. See, for example, J David Velleman, 'What Happens When Someone Acts?' *The Possibility of Practical Reason* (Oxford: Oxford University Press, 2000), 123-43.

he says, we can "learn about our own motivational states in the same way as we learn about the motivational states of others: that is, we infer them from behaviour."[22] It is clear that, when we are considering the attitudes of others, these inferences can go awry: we do not always correctly identify their beliefs and desires. Patten's claim is simply that we can make the same mistakes in our own case.

Patten argues that our beliefs can be influenced by what we expect to believe. This is why he denies that self-deception is necessarily motivated: what I *expect* to believe need not be a function of my desires.[23] When I am inferring my own motives from the evidence of my behaviour, my conclusions will reflect my beliefs about the sorts of motives *others* have when they behave similarly. But these beliefs may be mistaken—I might not understand others well at all—and even when they are right, my own behaviour might be the product of unorthodox motives. In both cases, I will be mistaken about my own motives. Patten's claim is that such mistakes need not be motivated by any desire: I may form the exact same beliefs about my motives as an unbiased third party.[24] Nevertheless, in being mistaken about what I believe, I will be self-deceived.

III. Self-deception and evidence

In Section V, I will argue that Patten's argument does not establish what he thinks it does. For now, I will only point out that, while Patten disagrees with both the intentionalists

[22] Patten, 'How Do We Deceive Ourselves?' at 231.
[23] Ibid.
[24] Ibid. at 233. To be clear, Patten is not denying that self-deception *can* be motivated, but only that such motivation is *necessary*. On his view, self-deception should not be defined as motivated false belief, even if many cases of self-deception are in fact motivated.

and the deflationists about the necessity of motivation for self-deception, all three camps share a basic agreement about the mechanisms by which self-deception proceeds. When we are self-deceived, it is because we interpret and gather evidence in a biased way. The disagreement is about whether such biases are the result of a desire, and, if so, whether they are intentional. Common to all parties in the debate, however, is an agreement the central role that mishandled evidence plays in self-deception. As we shall see, this has important consequences for the plausibility of the claim that we can be self-deceived abut our own attitudes.

Intentionalists are clear that self-deception involves the manipulation of evidence. Davidson even takes the fact the self-deceived agent handles evidence in a way that promotes the belief he wants to be true as support for his claim that this self-deception is intentional. He suggests that the self-deceived agent acquires his false belief by "obtaining new evidence in favour of believing" what he desires to be true, and by "pushing the negative evidence into the background or accentuating the positive."[25] The self-deceived agent is irrational because he does not have sufficient warrant for his belief, but he nevertheless manipulates the evidence to find some reason for his belief.[26]

Pears offers a similar story. The deceiving sub-agent forms beliefs in a way that parallels ordinary deliberation, and so generates motivated beliefs by a complex process of neutralizing evidence and avoiding confrontation with it. In fact, it is the apparent complexity of that deliberative process, and the way in which it seems to mirror ordinary

[25] Davidson, 'Deception and Division', at 209.
[26] In this respect, Davidson's account of self-deception mirrors his account of akrasia. The akratic agent has a reason for action; it is just not a *sufficient* reason, according to the agent's own judgment. Likewise, the self-deceived agent has evidence for his false belief; it is just not sufficient evidence.

deliberation and belief formation, that leads Pears to the view that self-deception must be intentional. It is "incredible", he says, that a tropistic mechanism such as the one suggested by Johnston could carry out the kind of complex information processing required to sustain self-deception over a long period of time, and yet we know that people can be self-deceived about their true characters for years, even decades.[27]

Mele and Johnston share much the same view of the role that the selective gathering and interpretation of evidence plays in self-deception; they simply deny that the process is intentional. Mele's explanation relies on a model for how we test hypotheses that claims that we generally do so in a way designed to minimize costly errors, rather than simply to track the truth.[28] If believing 'p' when p is false would be much more costly than believing 'not p' when p is true, then the hypotheses we consider, the evidence we collect, and the thresholds of proof we accept will be shaped so as to minimize the chance of falsely believing 'p.' Desires can cloud our evaluation of evidence, leading us to fail to count as relevant data that we should see as important, or even to count it as relevant in the wrong way.[29] They can also lead us to selectively focus our attention, concentrating on evidence that supports our point of view at the expense of data that contradicts it, and desires can affect the ways in which we gather evidence, leading us to be more sensitive to data that conforms to them than to data that contradicts

[27] David Pears, 'Self-Deceptive Belief Formation', *Synthese*, 89 (1991), 393-405 at 398.
[28] Mele's model is based on the "primary error detection and minimization" analysis of lay hypothesis testing defended in James Friedrich, 'Primary Error Detection and Minimization (Pedmin) Strategies in Social Cognition: A Reinterpretation of the Confirmation Bias Phenomenon', *Psychological Review*, 100 (1993), 298-319. Cited in Mele, *Self-Deception Unmasked*. Chapter 2-3.
[29] Mele's example is of a young man who takes the refusal of the object of his affections to talk to him as evidence that she is 'playing hard to get' and wants him to continue to pursue her. Mele, *Self-Deception Unmasked* at 26.

them.[30] Finally, there is the confirmation bias—the result of the fact that people testing a hypothesis tend to search for, and recognize, confirming instances more often than disconfirming ones. Since which hypotheses we test can be influenced by our desire, the confirmation bias can lead to motivationally biased false beliefs, and so to self-deception.[31] Mele describes much the same process as Davidson—the disagreement is simply over whether that process is launched by an intention.

The important point for our purposes is that each if the mechanisms identified by Mele involves failures in either the collection or the interpretation of evidence. It is by manipulating these cognitive processes that our desires lead us to—unintentionally—adopt biased false beliefs. Johnston identifies self-deception with similar strategies. In order to reduce anxiety, the mental tropism may lead one to: "Selectively reappraise and explain away the evidence (rationalization)… [A]void thinking about the touchy subject (evasion)… [F]ocus one's attention on invented reasons for p and spring to the advocacy of p whenever the opportunity presents itself (overcompensation)."[32]

Each of the strategies Johnston identifies—rationalization, evasion, and overcompensation—involves manipulating the gathering and interpretation of evidence in much the same way as the strategies identified my Mele, Pears, and Davidson.

Patten might disagree about the role of desire in self-deception, but he is in complete agreement about the role of misinterpreted evidence in such deception. On his view, we sometimes infer our own attitudes from the evidence supplied by our behaviour, and these inferences can go awry in various ways. His emphasis is that these mistakes are

[30] Ibid. at 26-8.
[31] Ibid. at 29.
[32] Johnston, 'Self-Deception and the Nature of Mind' at 75.

often a result of our expectations: if we believe that certain attitudes are generally associated with certain behaviour, we will assign those attitudes to ourselves when we exhibit those behaviours. But these inferences can be false, both because we can be wrong about which attitudes are generally associated with which behaviours, and because we can be mistaken about whether the generalizations on which we rely are true of *us*. Both of these mistakes involve flawed handling of the evidence, even if such mistakes need not be motivated by a desire.

All three views of self-deception—the intentionalist, the deflationist and Patten's non-motivational view—therefore agree that it arises because of failures in the gathering and interpreting of evidence. If this is right, then, as we shall see, an important consequence follows: despite Patten's claims, it is not possible for us to be self-deceived about the content of our own attitudes. That is because our knowledge of the content of these attitudes is not formed on the basis of evidence.

IV. First-person practical authority

The content of my own attitudes is something that I am uniquely placed to appreciate. Others can certainly learn what I believe, but they must do so by drawing inferences from the evidence provided by my words and deeds. Such inferences are generally reliable, though the possibility of deception—both intentional and otherwise—reminds us that they are far from infallible. I, on the other hand, do not form my beliefs about my own attitudes on the basis of evidence from my own behaviour. I have no need to consult such evidence, and no need to base my beliefs on inference from observation. My own

attitudes are among the things that, in Anscombe's phrase, I "know without observation."[33]

That I can know the content of my own mental life "without observation" is a reflection of a deep asymmetry between the first- and third-person perspectives: I have a special authority over the content of my own attitudes that others do not. As Davidson puts it, "first-person present-tense claims about thoughts, while neither infallible nor incorrigible, have an authority no second or third person claim, or first person other tense claim, can have."[34] Part of this authority concerns my immediate *access* to my own attitudes: as Richard Moran puts this point, I do not make judgments about the content of my attitudes based on inferences from anything epistemically more basic.[35] Others, by contrast, base their judgments about my attitudes on the evidence of my behaviour.

Davidson points out that to *identify* the special authority of the first-person perspective is not to explain it.[36] After all, the claim that we have special access to our own beliefs sounds suspiciously like an appeal to a thoroughly repudiated Cartesian view of the mind as a private "internal theatre" from which others are permanently barred entry. In fact, this Cartesian model actually undersells the difference between my access

---

[33] Anscombe, *Intention* at §8. Anscombe initially introduced the idea of non-observational knowledge as a way of talking about our knowledge of our own intentional actions, and this is still how the concept is often discussed: see, for example, Kieran Setiya, 'Practical Knowledge', *Ethics,* 118 (2008), 388-409, and Hanna Pickard, 'Knowledge of Action without Observation', *Proceedings of the Aristotelian Society,* 104 (2004), 205-30. The idea of non-observational knowledge applies equally well to our practical attitudes, however, since the point of both claims is that such knowledge is non-inferential.

[34] Donald Davidson, 'Knowing One's Own Mind', *Proceedings and Addresses of the American Philosophical Association,* 60 (1987), 441-58 at 441.

[35] Richard Moran, *Authority and Estrangement: An Essay on Self-Knowledge* (Princeton: Princeton University Press, 2001) at 10.

[36] Davidson, 'Knowing One's Own Mind', at 441.

to my own attitudes and the access others have to those same attitudes. To conceive of the mind as an "internal theatre" is see self-knowledge as a kind of inner perception. But that is to treat the question of the content of my attitudes as an empirical question, settled by a special kind of observation only I can undertake.

To say that I know the content of my attitudes "without observation," however, is to dispense with the idea that the question of what I believe is an empirical question, settled by appeal to the evidence within the theatre of the mind. The asymmetry between the first- and third-personal perspectives on the content of my beliefs is not merely that I have direct access to my beliefs while others have only indirect access. It is also that I stand in a completely different relation to the content of those beliefs that others do. My beliefs about my attitudes are different because those attitudes are *mine*. I determine what they are. This is part of what it means to say that I am a rational agent. As Richard Moran puts the point, "being a person whose mental life is brought to self-consciousness involves a stance of agency beyond that of being a kind of expert witness."[37]

To treat the content of my own attitudes as an empirical question would make me into a mere observer of my own mental life. But I do not take the content of my attitudes as a settled empirical question, because I do not simply observe my own attitudes: I determine them. The question of what I believe is therefore a *practical* one.[38] When I address the question of what I believe, I address at the same time the question of what I

[37] Moran, *Authority and Estrangement: An Essay on Self-Knowledge* at 4.
[38] As Moran puts it, I address the question "in a deliberative spirit." Of course, as I argued in Chapter 1, we do not generally form our beliefs on the basis of conscious deliberation. Moran's point is that there is "logical room" for the deliberative question of "what am I to believe.' He explicitly denies that this means that all of our beliefs are arrived at through a process of conscious deliberation. Ibid. at 63.

should believe, what I have most reason to believe. I do not determine what I believe by looking inward, to the stage of my internal theatre, but rather by turning my attention outward, away from me and towards what the world gives me reason to believe.

The same general point holds true for other attitudes, such as desires and intentions: when I address the question of what I want, I am generally at the same time addressing the question of what I have good reason to want.[39] Similarly, the question of what I intend is, for me, a question of that I have most reason to do. The first-person perspective on our own attitudes is therefore quite different from the third-person perspective. From the first-personal perspective, the question of what I believe, desire, or intend always has an essentially normative or practical dimension.[40] From the third-person perspective, however, this normative character is absent, and the question is entirely empirical. We can mark this distinction by referring to my current attitudes, whose content presents to me as a practical question, as 'practical attitudes.'

This claim should not be given an overly intellectualized or deliberative reading, particularly in light of the arguments I advanced in Chapter 2 against an intellectualized or deliberative understanding of intentional action. Each time I consciously entertain a belief, I do not rehash the reasons I have for holding it, and when I am moved by a desire,

---

[39] In other words, desire is a judgment sensitive attitude. Scanlon does not claim that *all* desires are judgment sensitive—thirst, for example, is not—only that *many* of them are in fact motivated by reasons. Scanlon, *What We Owe to Each Other* at 22. The distinction between motivated and unmotivated desires originates with Thomas Nagel, *The Possibility of Altruism* (Princeton: Princeton University Press, 1970) at 28-32.

[40] Moran emphasizes that this sort of self-knowledge is *practical*. Akeel Bilgrami makes a similar argument, but emphasizes the normative character of self-knowledge. Akeel Bilgrami, *Self-Knowledge and Resentment* (Cambridge, MA: Harvard University Press, 2006). Both contrast this sort of first-personal knowledge with the kind of empirical knowledge we have from the third-person perspective.

I do not first pause to consider the question of whether my desire is justified. In knowing what I believe and desire "without observation", I often know it without deliberation as well. But when what I believe is indeed a *question* for me, something for which I need to provide an answer, I do not settle it in an empirical mode, by observing myself. Rather, I settle it in the practical mode, by deciding what I have reason to believe.

The fact that we address the question of the content of our practical attitudes from a distinctly first-personal perspective is important, since all accounts of self-deception agree that it occurs because the self-deceived agent's desires (or sometimes, for Patten, his expectations) cause him to misinterpret evidence and so acquire a false belief in the face of that evidence. When we consider the content of our attitudes from the first-personal perspective, however, we do not do so by considering evidence, and so there is no evidence for us to misinterpret. Others may form false beliefs about our attitudes by misinterpreting evidence, but we may not. Our beliefs about the content of our own attitudes are practical, not evidential; they are expressions of agency, not the result of inference from empirical observation. This means that the mechanisms of self-deception can get no purchase on our beliefs about the content of our own attitudes, since those mechanisms all involve failures in the gathering and interpretation of evidence. Since our beliefs about the content of our own attitudes are non-evidential, *we cannot be self-deceived about the content of our own attitudes.*

V. How we can be wrong about ourselves

If the argument thus far is correct, then there is an important asymmetry between the first-person practical perspective we take on the content of our own attitudes, and the

third-person empirical perspective others take on those same attitudes. Our knowledge of our own beliefs is non-inferential, because our beliefs are under our own authority. The knowledge of others about our beliefs is inferential, since it is based on the evidence of our words and deeds. This asymmetry between the practical and empirical perspectives on our attitudes is radical, but its scope is limited.

To be clear, I am not intending to claim that we have infallible access to the content of our own minds, or that we are completely transparent to ourselves, or that we can never be estranged from our own attitudes.[41] Indeed, a central element of my agency is my ability to decide which of my desires I am going to embrace and which I wish to disavow.[42] Many facts about ourselves, including our own mental lives and intentional attitudes, are often mysterious, opaque, and we can be deeply mistaken about ourselves. My argument thus far has simply been that there is one particular kind of mistake—self-deception about the content of our own intentional attitudes—that we simply cannot make. But this does not mean that we are not obscure even to ourselves. In the space that remains, I want to consider some of the ways in which we can be mistaken about our own mental lives, and in so doing draw out some of the ethical implications of self-knowledge and its absence.

[41] Moran notes that the claim that there is an asymmetry between my access to my attitudes and third-personal access to those attitudes is a claim about a difference in *mode* of access: it does not mean that my judgments are either certain or incorrigible. His book, after all, deals with both authority and *estrangement.* If by judgments about my own attitudes go wrong, however, it will not be because of failures of inference from more basic evidence, as there is no such evidence and so my judgments about my attitudes are not based on such inferences.

[42] This is a central theme in much of Harry Frankfurt's work.

When I consider what someone *else* believes, I address this question from the third-person perspective. I treat the issue as an empirical one, and ask myself which inference best fits the evidence of his behaviour. But I must often consider *myself* from this same empirical perspective, and when I do, I can make mistakes and fall victim to self-deception. One of Davidson's examples of self-deception is Carlos, who believes that he will fail his driving test, and finds this belief painful. He therefore adopts the motivationally biased belief that he will *pass* his driving test.[43] Though Davidson's explanation of Carlos' self-deception is problematic in its reliance on the role of intention, the example is nevertheless instructive, since Carlos is self-deceived about *himself*. His mistake is that he fails to properly take up the third-person, empirical perspective on the question of his own driving ability, and assess it on the same grounds as his driving instructor.

Of course, Carlos' mistake was about his driving ability, not his own beliefs. But the asymmetry between the empirical and practical perspectives appears in relation to our own mental lives as well. When I consider what I *currently* believe, I address the question from the first-person, practical perspective, since I have authority over the content of my belief: it is up to me. But when I consider what I believed *in the past*, I no longer occupy the same practical perspective. I often, but not always, have better access to the content of my own past beliefs than others do, but that content is no longer up to me, and so there is no principled asymmetry between my own perspective on those beliefs and the perspective of a third party. For both of us, what I *used* to believe is an empirical question.

[43] Davidson, 'Deception and Division'.

Since I stand in an evidential rather than a practical relationship with the content of my past attitudes, I can be self-deceived about what I used to believe. My beliefs about by past attitudes depend on my memory. Memory, of course, is fallible. If I once believed something that is now embarrassing for me to admit, then pride can make it easier for me to forget that I ever believed it, and it can make it easier for me to think that I had the right belief all along.[44] If I can forget what I once believed, intended, cared about, and desired—and acute surprise and embarrassment at reading something one wrote many years ago is surely a common enough phenomenon—then I can be mistaken about the content of my own previous attitudes. When this forgetting is motivationally biased, as it so often seems to be, I am self-deceived about those same attitudes.[45]

It is the fact that we treat the content of our past attitudes as an empirical question that allows Patten to claim that we can be self-deceived about what we believe. His argument relies on the conclusions of several psychological experiments showing that, when asked to explain why we did something, our recall of our motives—and so of our attitudes—can be mistaken. These mistakes might be motivationally biased, though they can also arise simply because, as Patten points out, our beliefs are influenced by what we expect to believe. But even if we accept that these experiments provide evidence that we sometimes infer our attitudes from our behaviour, they do not show that we can be mistaken about what we currently believe. Rather, they show that we can be mistaken

[44] A process nicely captured by Nietzsche: "'I have done that,' says my memory. 'I cannot have done that' says my pride, and remains adamant. At last—memory yields." Friedrich Nietzsche, *Beyond Good and Evil*, trans. R.J. Hollingdale (London: Penguin Books, 1990) at §68.

[45] In fact, it is even possible to *intentionally* acquire false beliefs about the content of one's past attitudes, as I might do when I take a sleeping pill that introduces retroactive amnesia.

about what we believed *in the past*, since the experiments all involve subjects recalling their motives for past actions. Patten, however, puts his claim in the present tense: he argues that there can be a difference "between what an individual believes… and what he believes himself to believe."[46] This shift from the past to the present tense is illegitimate, as it overlooks the asymmetry between the first- and third-person perspectives on our own mental lives. All Patten's arguments can show is that we can be self-deceived about what we *used to* believe. He fails to show that we can be self-deceived about what we *currently* believe, because such self-deception is impossible.

The failures of self-knowledge that allow others to know us better than we know ourselves cannot extend as deep as our knowledge of the content of our own attitudes, but that does not mean that our self-knowledge with respect to our own mental lives is either perfect or infallible. I have just argued that we can be mistaken about the content of our own past beliefs. We can also be mistaken about our current beliefs, so long as those mistakes are not about the content of such beliefs. In fact, current beliefs are at the centre of the clearest example of *counterprivacy*; facts others can know about me that I cannot know about myself. As G.E. Moore pointed out, there is an absurdity in the claim "I believe that it is raining, but it is not raining." To believe that p just is to believe that p *is true.* But though the claim is absurd coming from my mouth, the truth-functionally equivalent "he believes that p, but p is false", can be truthfully and easily uttered by someone else. After all, I surely have many false beliefs, and others can know which of my beliefs are false, even if I cannot. In a similar vein, Socrates famously argued that I

[46] Patten, 'How Do We Deceive Ourselves?' at 230.

can never knowingly choose the bad, though others can know that I do.[47] In such cases, my second-order beliefs are mistakes in the *evaluation* of my attitudes as true, or good, rather than mistakes about the content attitudes of those attitudes. I might know what I believe, but I do not always know how that belief ought to be assessed or evaluated. In other words, though it is not possible to be self-deceived about what it is that we currently believe, it may be possible for us to be self-deceived about the proper assessment of our own beliefs.

In fact, mistaken evaluations of our current attitudes can easily be the product of self-deception. Emma, for example, is self-deceived about her matchmaking ability. At a deeper level, though, she is self-deceived about the status of her own beliefs: she thinks that they are dispassionate and objective, when in fact they are the product of her desire to be a matchmaker; she thinks they are true, when they are false; she thinks they are justified by the evidence, when in fact a genuinely skilful observer of human nature would have noticed that Mr. Elton was attracted to *her*, not Harriet. Like many other victims of self-deception, Emma fails to appreciate the ways in which her attitudes are biased by her desires, and so fails to clearly see herself as self-deceived. In Holton's terms, Emma is "mistaken about whether [she] is living up to [her] own belief-forming standards."[48] The hypocritical environmentalist is similar: she thinks that her beliefs are true although they are false, but she *also* believes that her beliefs are the result of a

[47] Plato, *Protagoras*. For an argument that we *can* knowingly choose the bad, see, for example, Michael Stocker, 'Desiring the Bad: An Essay in Moral Psychology', *The Journal of Philosophy*, 76 (1979), 738-53. For a discussion of counterprivacy that suggests that Moore identified more than one paradox, and that connects the case of belief to the Socratic case of practical reason, see André Gombay, 'Some Paradoxes of Counterprivacy', *Philosophy*, 63 (1988), 191-210.

[48] Holton, 'What Is the Role of the Self in Self-Deception?' at 60.

dispassionate evaluation of the evidence, rather than a biased attempt to see herself in a positive light, and so she thinks that her actions are principally motivated by a concern for the environment, rather than a love of her own self-image.

Emma, like the hypocrite, is therefore making two mistakes. She is wrong about the status of her beliefs as true, objective, and justified, and she is therefore also wrong about *herself*: she thinks that she is a sympathetic and objective observer of others as well as a skilled matchmaker. Both mistakes concern her attitudes, though in different ways: the first mistake is about the evaluation of her particular beliefs, and the second is a more global mistake about her disposition to form correct beliefs.

The various kinds of mistakes outlined above can be mutually reinforcing. My pride can lead me to edit my memory of my past beliefs, and self-deception about the content of my previous beliefs will often involve believing that I *used* to believe what I *now* take to be the truth. This, in turn, can increase my confidence that my *current* beliefs are true, objective, and justified. After all, I might reason, I have been right in the past, so I am likely to be right in the present. And both of these mistakes can lead me to suppose that I am the sort of person who always gets things right. In the other direction, my confidence in my own cognitive abilities can increase my confidence in my exercise of those abilities in the *past*, and so lead me to falsely conclude that I used to believe what I now take to be the truth. In general, mistakes about the content of my past attitudes can lead to mistaken self-assessments in the present, and such mistaken self-assessments can, in turn, colour the ways in which I look back on the past. The mistaken self-assessments that emerge from self-deception are woven very fine, and can be mutually reinforcing.

The possibility of such pervasive and mutually reinforcing self-deception means that a lack of self-knowledge can extend deep into our own mental lives. We can be mistaken about what we have believed, about the proper assessment of our beliefs, both past and present, and, more broadly, about how we are disposed with regard to the formation of beliefs. We might think that we are disposed to always form true, objective, and justified beliefs, and yet be mistaken about that, just as Emma is mistaken about whether her beliefs about Mr. Elton are justified.

The same point holds for other attitudes: we may not be self-deceived about the content of our current desires, but we can be self-deceived about what we desired in the past, about what we are likely to desire in the future, and about how we ought to assess our desires, past, present, and future. We can be wrong about *why* we desire it, just as Emma is wrong about why she believes Mr. Elton loves Harriet. We can be wrong about whether our desires are ethically justified, and about whether what we desire is good, and worth desiring, or not. Again, more broadly, such mistakes can lead to more global failures of self-knowledge, if we become mistaken about what and how we are *disposed* to desire, and about whether those dispositions are admirable or blameworthy.

If this is right, then our deepest failures of self-knowledge are not about what we believe or desire: they are about our attitudinal dispositions. In other words, when others truly know us better than we know ourselves, what they know, and we do not, is the quality of our characters, the kind of people that we truly are. Emma is wrong in thinking that she is a skilled matchmaker and sympathetic observer of the human scene; Huck is mistaken in thinking that he is a weak boy who acts for bad reasons and is beyond redeeming; the hypocrite is wrong to think that she is a paragon of virtue. These are, in a

sense, failures of self-knowledge similar to Carlos' self-deception about his driving ability: they are about mistakes about the possession of skills, capacities, and dispositions, rather than the possession of particular attitudes. But, unlike Carlos' self-deception, they are mistakes about the possession of capacities and dispositions that are essentially related to attitudes: the capacity and disposition to form true and justified beliefs, to act for good reasons, and to have justified desires.

In one sense, a failure to accurately assess one's own character is not as deep as a failure to know the content of one's own attitudes. After all, no matter how profound one's failures of self-knowledge, self-deception cannot touch the authority one has over the content of one' own attitudes. In another sense, however, this failure is even deeper. I may I know what I believe and desire and yet be deeply mistaken about the origin and justifiability of those attitudes, about the content and status of my attitudes in the past, and about the attitudes I am likely to have in the future. As my self-deception increases, my knowledge of my own attitudes becomes increasingly disconnected from other areas of self-knowledge. So while my first-person rational authority over my own attitudes remains unthreatened by self-deception, the importance of that authority diminishes as my lack of self-knowledge of my own character increases. We cannot be self-deceived about what we think, but we can be self-deceived about the quality of our own characters, and this mistake is much more profound

VI. Conclusion: overcoming our mistakes

Over the course of the last five chapters, I have been arguing that failures of self-knowledge are morally significant. A lack of self-knowledge can lead directly to other,

serious moral failings. Those who get things wrong morally speaking often do so because they get things wrong about *themselves.* Such failures of self-knowledge can lead to a failure to extend one's moral sensitivity, to an alienation from one's true moral judgments, to moral complacency, and to moral blindness. In other words, it can lead to the sorts of failures that beset the person who only occasionally attains the deliberative outlook of virtue, the inverse akratic, and the hypocrite.

One important consequence of this view is that moral education and moral development unavoidably involves a search for self-knowledge. This does not mean that self-knowledge, on its own, is sufficient for virtue, both because it is likely possible for someone be both self-aware and unmoved by morality, and because moral education involves more than *just* learning about oneself. Moral education involves learning which actions are right, and why; it involves coming to be moved to do act in morally desirable ways, for the right reasons and with the right deliberative outlook; and it involves learning which ends and projects are truly desirable and worthwhile, and coming to desire and to take pleasure in just those ends and projects. But unless this sort of moral knowledge is complemented by an equivalent investment in self-knowledge, moral development is bound to be at best incomplete. So it turns out that Socrates was right, and 'know thyself' really is a fundamental ethical imperative.

There is a worry, however, about this emphasis on the moral importance of self-knowledge. It might appear that, in highlighting the close connection between self-knowledge and virtue, I am advocating for an objectionably intellectualized and self-absorbed understanding of virtue. That is, it might appear that virtue involves, not simply the possession of practical wisdom, but a particularly reflexive and self-absorbed form of

concern. But the idea of the virtuous person as someone 'caught up in their own head,' engaged in constant conscious self-monitoring and self-assessment, is not an attractive one. Indeed, as I emphasized in discussing hypocrisy and self-deception, preoccupation with one's own moral status is often a sign of misplaced ethical concern, rather than a mark of virtue. But the claim that self-knowledge is a fundamental ethical imperative need not be given an overly intellectualized reading. What is important is not primarily the constant self-assessment and the conscious formulation of a detailed and accurate self-description. Rather, what matters, with regard to self-knowledge, is avoiding mistaken self-conceptions; that is to say, what matters is avoiding allowing the mistaken self-assessments that can flow from a love of "the dear self" to cloud one's deliberative outlook in ways that are liable to lead to error. The self-conception of virtue need not be something of which the virtuous agent is always aware: it should often be like a clear pane of glass, allowing one to transparently and accurately see the how one ought to live.

Some might object that the right conclusion to draw from the close connection between moral formation and self-knowledge is not that we have a moral obligation to gain self-knowledge, but rather that those whose apparent moral failings emerge from mistaken self-assessments should be excused from blame. After all, I can be ignorant about myself without that ignorance being blameworthy: If I hit my head, I am not expected to know if I have a concussion, and I can leave it to the doctor to diagnose me. Why should I be responsible when my ignorance is about the quality of my character rather than the nature of my headaches? If Emma is not at fault for being wrong about her own character, the argument might go, then we should not blame her for the apparent flaws that arise from her error. This suggestion is particularly attractive if one accepts

that self-deception is not intentional: if she does not intend to deceive herself, then why should we blame her so harshly for her mistake?[49] Perhaps, then, the real moral importance of self-deception is that we ought to excuse far more people than we currently do. If their apparent failings come from a mistake for which they are not responsible, then they should be excused from blame.

I would appear to have good reason to embrace this line of argument. After all, I have been maintaining that imperfect agents are often treated too harshly, and so any argument that would allow me to let the self-deceived off the hook would be more support for my position. Despite its superficial attractions, however, I want to resist this argument.

The lack of self-knowledge displayed by the person with limited moral sensitivity, by Huck and other inverse akratics, by hypocrites like our righteous environmentalist and Orwell's humanitarian, and by Emma and others who are self-deceived, are not just mistakes. They are *failures.* Such characters are not simply ignorant; they are at fault for their ignorance. This ignorance can be the result of a lack of effort, of a too-strong concern for their own self-image, or of complacency in the face of the demands of morality. Each of these mistakes can be blameworthy.

A consequence of our practical attitudes being an expression of our rational agency is that they are, in an important sense, 'up to us': extraneous influences such as unacknowledged desires and prejudices may play a role in the formation of our beliefs, but *we* play a central role as well. Though we may not intend to form false beliefs about

[49] Neil Levy argues that if self-deception is indeed non-intentional then it is generally a mistake to hold the self-deceived responsible for their condition. Neil Levy, 'Self-Deception and Moral Responsibility', *Ratio,* XVII (2004), 294-311.

ourselves, we can nevertheless correct those false beliefs, and get a handle on our own attitudes and dispositions. We can try harder to set aside the influence of desire, to root out the prejudices that cloud our thinking, and to ensure that our beliefs are formed on the basis of all of the evidence.

If this sort of self-improvement is possible, it will require making an effort to attain an honest and accurate self-understanding. Failing to undertake this effort can render us the legitimate targets of blame. But we do not need to undertake this effort alone. One of the consequences of the fact that others can know us better than we know ourselves is that they can help us understand ourselves. When they have insights and criticisms of our actions and our characters, we have an obligation to listen.

Bibliography

ANSCOMBE, G.E.M. (1957), *Intention* (Cambridge, MA: Harvard University Press).

--- (1981), 'Modern Moral Philosophy', *Collected Philosophical Papers Volume III: Ethics, Religion, Politics* (Minneapolis: University of Minnesota Press).

ARISTOTLE (1999), *Nicomachean Ethics*, trans. Terence Irwin (2nd edn.; Indianapolis: Hackett).

ARPALY, NOMY (2000), 'On Acting Rationally against One's Best Judgment', *Ethics*, 110, 488-513.

--- (2003), *Unprincipled Virtue* (Oxford: Oxford University Press).

--- (2006), *Merit, Meaning, and Human Bondage: An Essay on Free Will* (Princeton: Princeton University Press).

AUSTEN, JANE (1981), *Emma* (New York: Bantam Classic).

BADHWAR, NEERA (1996), 'The Limited Unity of Virtue', *Noûs*, 30, 306-32.

BENNETT, JONATHAN (1974), 'The Conscience of Huckleberry Finn', *Philosophy*, 49, 123-34.

BILGRAMI, AKEEL (2006), *Self-Knowledge and Resentment* (Cambridge, MA: Harvard University Press).

BOKSEM, MAARTEN, MEIJMAN, THEO, AND LORIST, MONICQUE (2005), 'Effects of mental fatigue on attention: An ERP study', *Cognitive Brain Research*, 25, 107-16.

CRISP, ROGER AND COWTON, CHRISTOPHER (1994), 'Hypocrisy and moral seriousness', *American Philosophical Quarterly*, 31, 343-8.

CROMBEZ, GEERT, et al. (1996), 'The Disruptive nature of Pain: An Experimental Investigtion', *Behav. Res. Ther.*, 34, 911-18.

DAVIDSON, DONALD (1980a), 'Freedom to Act', *Essays on Actions and Events* (Oxford: Oxford University Press).

--- (1980b), 'Actions, Reasons, and Causes', *Essays on Actions and Events* (Oxford: Oxford University Press).

--- (1984), 'Thought and Talk', *Inquiries into Truth and Interpretation* (Oxford: Clarendon Press), 155-70.

--- (1987), 'Knowing One's Own Mind', *Proceedings and Addresses of the American Philosophical Association*, 60, 441-58.

--- (2001), 'How is Weakness of the Will Possible?' *Essays on Actions and Events* (Oxford: Clarendon Press).

--- (2004), 'Deception and Division', *Problems of Rationality* (Oxford: Clarendon Press), 199-212.

DE SOUSA, RONALD (2002), 'Emotional Truth', *Proceedings of the Aristotelian Society: Supplementary Volume*, 76, 247-63.

DORIS, JOHN (2002), *Lack of Character: Personality and Moral Behaviour* (Cambridge: Cambridge University Press).

DRIVER, JULIA (1996), 'The Virtues and Human Nature', in Roger Crisp (ed.), *How Should One Live? Essays on the Virtues* (Oxford: Clarendon Press), 111-30.

FARRIN, LYDIA, et al. (2003), 'Effects of Depressed Mood on Objective and Subjective Measures of Attention', *Journal of Neuropsychiatry and Clinical Neuroscience*, 15, 98-104.

FEDER KITTAY, EVA (1982), 'On Hypocrisy', *Metaphilosophy,* 13, 277-89.
FOOT, PHILIPPA (2002), 'Rationality and Virtue', *Moral Dilemmas* (Oxford: Clarendon Press), 159-74.
FRANKFURT, HARRY (1988a), 'Freedom of the will and the concept of a person', *The Importance of What We Care About* (Cambridge: Cambridge University Press).
--- (1988b), 'Rationality and the Unthinkable', *The Importance of What We Care About* (Cambridge: Cambridge University Press), 177-90.
--- (1988c), *The Importance of What We Care About* (Cambridge: Cambridge university Press).
FRIEDRICH, JAMES (1993), 'Primary Error Detection and Minimization (PEDMIN) Strategies in Social Cognition: A Reinterpretation of the Confirmation Bias Phenomenon', *Psychological Review,* 100, 298-319.
FULLINWIDER, ROBERT (2005), 'On Moralism', *Journal of Applied Philosophy,* 22, 105-20.
GOMBAY, ANDRÉ (1988), 'Some Paradoxes of Counterprivacy', *Philosophy,* 63, 191-210.
GREENSPAN, PATRICIA (1988), *Emotions and Reasons: An Inquiry into Emotional Justification* (London: Routledge).
HARMAN, GILBERT (1999), 'Moral Philosophy Meets Social Psychology: Virute Ethics and the Fundamental Attribution Error', *Proceedings of the Aristotelian Society,* 99, 315-31.
HEIL, JOHN (1984), 'Doxastic incontinence', *Mind,* 93, 56-70.
HOLTON, RICHARD (1999), 'Intention and Weakness of the Will', *The Journal of Philosophy,* 95, 241-62.
--- (2001), 'What is the Role of the Self in Self-Deception?' *Proceedings of the Aristotelian Society,* 101, 53-69.
HUME, DAVID (1757) 'The Natural History of Religion'.
HURKA, THOMAS (2006), 'Virtuous act, virtuous dispositions', *Analysis,* 66, 69-76.
HURSTHOUSE, ROSALIND (1999), *On Virtue Ethics* (Oxford: Oxford University Press).
ISHIGURO, KAZUO (1988), *The Remains of the Day* (New York: Vintage International).
JOHNSTON, MARK (1988), 'Self-Deception and the Nature of Mind', in Brian McLaughlin and A.O. Rorty (eds.), *Perspectives on Self-Deception* (Berkeley: University of California Press), 63-91.
KANT, IMMANUEL (1993a), 'On a Supposed Right to Lie because of Philanthropic Concerns', *Groundwork for the Metaphysics of Morals* (Indianapolis Hackett Publishing).
--- (1993b), *Groundwork for the Metaphysics of Morals*, trans. James W Ellington (3rd edn.; Indianapolis: Hackett).
LEVY, NEIL (2004), 'Self-Deception and Moral Responsibility', *Ratio,* XVII, 294-311.
MACKINNON, CHRISTINE (1991), 'Hypocrisy, with a note in integrity', *American Philosophical Quarterly,* 28 (4), 321-30.
--- (2002), 'Hypocrisy and the Good of Character Possession', *Dialogue,* XLI, 715-39.
MCDOWELL, JOHN (1981), 'Non-cognitivism and rule-following', in C Leich and S Holtzman (eds.), *Wittgenstein: To Follow a Rule* (London: Routledge).
--- (1997), 'Virtue and Reason', in Roger Crisp and Michael Slote (eds.), *Virtue Ethics* (Oxford: Oxford Univrtsity Press), 141-62.

MCINTYRE, ALISON (1990), 'Is Akratic Action Always Irrational?' in Owen Flanagan and A.O. Rorty (eds.), *Identity, Character, Morality: Essays in Moral Psychology* (Cambridge, MA: MIT Press), 380-400.
MELE, ALFRED (1987), *Irrationality* (Oxford: Oxford University Press).
--- (2000), *Self-Deception Unmasked* (Princeton: Princeton University Press).
MILL, JOHN STUART (2001), *Utilitarianism* (Indianapolis: Hackett).
MORAN, RICHARD (2001), *Authority and Estrangement: An Essay on Self-Knowledge* (Princeton: Princeton University Press).
--- (2005), 'Problems of Sincerity', *Proceedings of the Aristotelian Society*, 105, 341-61.
NAGEL, THOMAS (1970), *The Possibility of Altruism* (Princeton: Princeton University Press).
NIETZSCHE, FRIEDRICH (1990), *Beyond Good and Evil*, trans. R.J. Hollingdale (London: Penguin Books).
ORWELL, GEORGE (1970), 'Rudyard Kipling', *The Collected Essays, Journalism, and Letters of George Orwell, Volume 2: My Country Right or Left* (Harmondsworth: Penguin), 215-29.
PATTEN, DAVID (2003), 'How do we deceive ourselves?' *Philosophical Psychology*, 16, 229-46.
PEARS, DAVID (1984), *Motivated Irrationality* (Oxford: Oxford University Press).
--- (1991), 'Self-Deceptive Belief Formation', *Synthese*, 89, 393-405.
PENNER, TERRY (1973), 'The Unity of Virtue', *The Philosophical Review*, (82), 35-68.
PICKARD, HANNA (2004), 'Knowledge of Action Without Observation', *Proceedings of the Aristotelian Society*, 104, 205-30.
PLATO (1996), *Protagoras*, trans. C.C.W. Taylor (Oxford: Oxford University Press).
RAZ, JOSEPH (1990), *Practical Reason and Norms* (2nd edn.; Oxford: Oxford University Press).
--- (1999), *Engaging Reason* (Oxford: Oxford University Press).
ROBINSON, JENEFER (1995), 'Startle', *The Journal of Philosophy*, 92, 53-74.
SCANLON, T.M. (1998), *What We Owe to Each Other* (Cambridge, MA: Belknap Press).
--- (2002), 'Reasons and Passions', in Sarah Buss and Lee Overton (eds.), *Contours of Agency: Essays on Themes from Harry Frankfurt* (Boston: MIT Press), 165-83.
--- (2008), *Moral Dimensions: Permissibility, Meaning, Blame* (Cambridge, MA: Harvard University Press).
SETIYA, KIERAN (2007), *Reasons Without Rationalism* (Princeton: Princeton University Press).
--- (2008), 'Practical Knowledge', *Ethics*, 118, 388-409.
SHAW, WILLIAM (1999), 'Is Hypocrisy a Problem for Consequentialism?' *Utilitas*, 11, 340-46.
SHKLAR, JUDITH (1984), 'Let us not be hypocritical', *Ordinary Vices* (Cambridge, MA: Belknap Press of Harvard University Press), 45-86.
SINGER, PETER (1981), *The Expanding Circle: Ethics and Sociobiology* (Oxford: Oxford University Press).
SMITH, ANGELA (2005), 'Responsibility for Attitudes: Activity and Passivity in Mental Life', *Ethics*, 115, 236-71.
SOIFER, ELDON AND SZABADON, BELA (1998), 'Hypocrisy and Consequentialism',

Utilitas, 10, 168-94.
--- (1999), 'Hypocrisy, Change of Mind, and Weakness of Will: How to do Moral Philosophy with Examples', *Metaphilosophy*, 30, 60-78.
--- (2004), *Hypocrisy: Ethical Investigations* (Peterborough: Broadview).
SREENIVASAN, GOPAL (2002), 'Errors about Errors: Virtue Theory and Trait Attribution', *Mind*, 111, 47-68.
STATMAN, DANIEL (1997), 'Hypocrisy and Self-Deception', *Philosophical Psychology*, 10, 57-78.
STOCKER, MICHAEL (1979), 'Desiring the Bad: An Essay in Moral Psychology', *The Journal of Philosophy*, 76, 738-53.
STRAWSON, P.F. (1982), 'Freedom and Resentment', in Gary Watson (ed.), *Free Will* (Oxford: Oxford University Press).
TURNER, DAN (1990), 'Hypocrisy', *Metaphilosophy*, 21, 262-69.
TWAIN, MARK (1994), *The Adventures of Huckleberry Finn* (New York: William Morrow).
VELLEMAN, J DAVID (2000), 'What Happens When Someone Acts?' *The Possibility of Practical Reason* (Oxford: Oxford University Press), 123-43.
WATSON, GARY (2004), 'Skepticism About Weakness of Will', *Agency and Answerability* (Oxford: Clarendon Press), 33-58.
WIGGINS, DAVID (1980), 'Weakness of Will, Commensurability, and Objects of Desire', in A.O. Rorty (ed.), *Essays on Aristotle's Ethics* (Berkeley: University of California Press).
WILLIAMS, BERNARD (1973), 'Deciding to Believe', *Problems of the Self* (Cambridge: Cambridge University Press), 136-51.
--- (1981), 'Persons, character and morality', *Moral Luck* (Cambridge: Cambridge University Press), 1-19.
--- (1985), *Ethics and the Limits of Philosophy* (Cambridge, MA: Harvard University Press).
--- (1995a), 'Voluntary acts and responsible agents', *Making Sense of Humanity* (Cambridge: Cambridge University Press), 22-34.
--- (1995b), 'Moral incapacity', *Making Sense of Humanity* (Cambridge: Cambridge University Press), 46-55.
WOLF, SUSAN (2007), 'Moral Psychology and the Unity of the Virtues', *Ratio*, XX, 145-67.