Desire and the Rationality of Virtue

by

Andrés Carlos Luco

Department of Philosophy
Duke University

Date:_______________________

Approved:

___________________________

Owen Flanagan, Co-Supervisor

___________________________

David Wong, Co-Supervisor

___________________________

Allen Buchanan

___________________________

Kevin Hoover

___________________________

Alexander Rosenberg

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Philosophy in the Graduate School
of Duke University

2009
ABSTRACT

Desire and the Rationality of Virtue

by

Andrés Carlos Luco

Department of Philosophy
Duke University

Date: _______________________

Approved: ____________________

Owen Flanagan, Co-Supervisor

David Wong, Co-Supervisor

Allen Buchanan

Kevin Hoover

Alexander Rosenberg

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Philosophy in the Graduate School of Duke University

2009
Abstract

A classic question in moral philosophy asks “Why be moral?” In other words, what reason or motive do people have to act in accordance with the requirements of morality? In the tradition of Thomas Hobbes and David Hume, this project defends the thesis that nearly all people have reason to be moral nearly all of the time, because moral conduct generally serves individuals’ desires and needs. It’s argued, first, that a reason for action must be capable of motivating an agent to act, and second, that reasons for action motivate through the desires of the agent. This view is defended against the objection that reasons for action are not contingent on any particular agent’s desires. Turning to morality, the case is made that the desires of an individual can be consonant with the demands of morality in any of three possible ways: (1) moral action serves one’s other-regarding desires to help others; (2) moral action serves one’s moral desires, which are formed when one internalizes the moral norms of his or her community; and (3) moral action serves one’s self-regarding desires to avoid punishments that one incurs by violating moral norms. In the final chapter, it is acknowledged that the moral norms which happen to prevail in a society sometimes conflict with the moral convictions of individuals. Under certain conditions, however, it can be rational for nearly all members of a society to collectively change existing moral norms. Furthermore, it is within the power of individuals to foment the conditions for collective transitions to alternative moral norms.
Dedication

To my family
Contents

Abstract .............................................................................................................................. iv

List of Figures ..................................................................................................................... x

Acknowledgements ............................................................................................................ xi

Introduction: Desire and the rationality of virtue ............................................................... 1

1. Chapter 1: Practical rationality and the internalism-externalism debate ...................... 13
   1.0 What is practical rationality? ................................................................................ 13
   1.1 Instrumentalism about reasons .......................................................................... 17
   1.2 Historical origins of instrumentalism ................................................................. 38
   1.3 Instrumentalism and internalism about reasons .................................................. 41
   1.4 Actual reasons internalism .................................................................................. 45
      1.4.1 Procedural vs. substantive rationality ............................................................. 45
      1.4.2 Instrumentalism as an actual internalism ....................................................... 49
      1.4.3 Williams’s actual internalism ...................................................................... 54
   1.5 Counterfactual reasons internalism: Practical reasoning beyond the motivational set? .................................................................................................................. 77
   1.6 Reasons Externalism: Practical reasoning beyond the explanatory constraint? ... 87
      1.6 Conclusion ........................................................................................................ 102

2. Chapter 2: Practical reasons and motivation ............................................................... 105
   2.0 Normative versus explanatory reasons for action ................................................. 105
   2.1 The master argument for Williams’s explanatory constraint ................................ 109
   2.2 “Reasons Imply Can” ....................................................................................... 118
   2.3 Acting for reasons ............................................................................................... 123
4.4.1 Balancing two intuitions: the subjectivity of reasons vs. the fallibility of ends .......................... 325

4.4.2 Choosing and pursuing ends ............................................................................................... 329

4.4.3 Reflective instrumentalism: leaving no loose ends .................................................. 336

4.4.4 Unification: the fallibility of ends and the subjectivity of reasons .................. 343

4.5 Conclusion .................................................................................................................. 349

5. Chapter 5: The rationality of moral compliance .......................................................... 351

5.0 Introduction ............................................................................................................. 351

5.1 Moral norms and being moral .................................................................................. 356

5.2 Moral norms and mixed-motive games .................................................................. 366

5.3 Moral norms make a difference ............................................................................. 375

5.4 The instrumental rationality of following norms ..................................................... 389

5.4.1 Punishment and rule egoism: Hobbes’s reply to the fool ................................ 390

5.4.2 Norm-conformity as a virtue: Hume’s response to the sensible knave........ 406

5.4.3 Overlapping reasons to be moral ........................................................................ 422

5.5 Unpopular norms: the dark side of conformity ...................................................... 426

5.6 Persuasive communication, “fundamental moral norms,” and basic human needs .................................................................................................................. 452

5.6.1 Internalization by persuasion .................................................................................. 452

5.6.2 Basic human needs as Sperberian biases .......................................................... 459

5.6.3 Abandonment of FGC: explaining the human rights effect ......................... 488

5.6.4 Moral compliance and moral dissidence .......................................................... 489

5.7 Conclusion .................................................................................................................. 491
List of Figures

Figure 1: The prisoner’s dilemma................................................................. 368
Figure 2: Norm-transformation of the prisoner’s dilemma in figure 1........... 413
Figure 3: A coordination game with indifferent equilibria.......................... 434
Figure 4: A coordination game with ranked equilibria................................. 441
Acknowledgements

This dissertation was completed with the support of a number of people. I am grateful to all of them. Of course, I profited immensely from the feedback of my dissertation committee—David Wong, Owen Flanagan, Allen Buchanan, Alex Rosenberg, and Kevin Hoover. Their comments on my work never failed to illuminate blind spots in my arguments. They also showed tremendous patience with me as I labored to finish and revise chapter drafts. My graduate student colleagues in the Department of Philosophy at Duke University offered a challenging, but always friendly intellectual environment in which to develop my ideas. For over three years, I was employed in the Language Labs of the Center for Instructional Technology (CIT) at Duke. In fact, I was hired as the lone graduate student on the Student Assistant staff. I give thanks to my employers in the CIT, who provided me with the opportunity to work part-time while allowing me the latitude I needed to carry out my research. The sometimes arduous day-to-day task of writing a dissertation would not have been nearly so bearable without the friendship of Beatriz Collada. Her generosity, kindness, and encouragement were as sure as the supply of Cuban pastries she would bring from Miami. Finally, there is simply no way to overstate the ways in which the completion of this dissertation was made possible by my family—Juan Enrique, Nina, and Nicolás Luco. Through the long trials of graduate school, they kept me focused on what’s most important, and they served as a constant, comforting reminder of where I came from.
Introduction: Desire and the rationality of virtue

Morality makes heavy demands of us. Among other things, it tells us how we ought to live, what goals we ought to pursue, and how to balance our own interests against those of others. Our desires and inclinations often lead us away from doing “the right thing.” And it’s hardly surprising that, whenever duty and desire are at odds with each other, people frequently just do what they want. In less developed countries, thousands of children die every year from malnutrition and disease. Yet people in wealthy countries spend far more money on beer and perfume than their governments devote to foreign aid. By giving up just a small amount of consumer satisfaction, we could do much to improve the lot of those who live in such desperate conditions. To some, this fact alone is sufficient to ground a moral obligation to help the world’s poor. But when it comes down to it, many of us don’t muster up the will to carry out the obligations that we recognize ourselves to have.

These considerations should raise doubts about the audacious pretensions of morality. After all, principles of morality, justice, virtue are commonly thought to be **overriding**. That is, they provide standards for evaluating conduct which are independent of any particular person’s desires or interests. A husband may find an extramarital affair so gratifying that he is willing to carry on with it as long as it suits him. Nevertheless, he has obligations to his wife and family that render what personal satisfaction he gets from his infidelity beside the point. A husband has a duty to be faithful to his wife, regardless of whether or not he **wants** to be. His duty overrides his desire, not the other
way around. Similarly, many of us think we have positive moral duties to help other people, provided that it does not require a sacrifice of comparable moral significance on our part. For instance, Peter Singer has argued that the citizens of rich nations should devote around 10 percent of their income to combat global poverty (Singer, 1993). Obviously many people would not be inclined to do that, but for Singer this is our duty nonetheless. We have a duty to help improve the condition of those in abject poverty, and this duty applies to us whether we want to fulfill it or not. Even if one does not think morality imposes such demanding positive duties, few would deny that we are, at the very least, subject to negative duties not to harm or deceive others. Yet every day we encounter people who insult, injure, lie, cheat, steal, or even kill for personal gain. Many people willingly carry out these deeds, but it remains the case that they acted against their duties.

If morality is overriding, you cannot escape your moral duty simply by declaring that you don’t want to do it. This brings into view a potentially unappealing feature of morality—namely, that living a moral life can be burdensome. Students cheat on tests, and spouses on each other, presumably because they get something they want out of it. It’s no mystery why people are attracted to things like passionate romances or perfect grade point averages. So why should anyone let morality stand in the way of getting what one wants? If untold satisfactions stand to be gained from defying moral requirements, then doing one’s duty can be a heavy burden indeed. That morality can be so burdensome
raises a philosophical problem for the ages: *Why be moral?* What reason do we have to lead ethical, just, virtuous lives?

The problem broached here has received several classic formulations in the history of philosophy. All of them echo the same basic point. Leading an ethical life can be burdensome. The burdens consist largely of losing opportunities for advantage that can be secured through unethical conduct. But why should we accept those burdens? If someone encounters a possibility for advantage, it seems obvious that she has every reason to pursue it. When the pursuit of advantage would be unethical, what reason for moral restraint could there be? Glaucon, brother of Plato, puts this challenge to Socrates in Book II of *Republic*.¹ He argues that most people believe an unjust life to be “more profitable” than a just one. People by nature desire “to outdo others and get more and more,” but considerations of justice often forbid this (Plato, *Republic II*, 359c). What prevents people from seizing the advantages of injustice are the fear of punishment by others, and the rewards that accompany a reputation for being just. But Glaucon implores us to imagine someone with the power to claim the advantages wrought by injustice without opening himself up to punishment or a besmirched reputation. Suppose a man could put on a magic ring that made him invisible to others, leaving him free to do whatever he pleased. Glaucon predicts that any such person would “take whatever he wanted from the market-place with impunity, go into people’s houses and have sex with

anyone he wished, kill or release from prison anyone he wished, and do all the other things that would make him like a god among humans” (Plato, *Republic II*, 360c). Glaucon is not merely making a point about the dark side of human psychology. He goes on to assert that anyone who refrains from injustice in the face of a perfect opportunity for advantage would be considered “wretched and stupid by everyone aware of the situation” (Plato, *Republic II*, 360d). Glaucon is presenting what he considers the popular view about what it makes sense for a person to do when he has the chance to benefit from injustice. On this view, any advantage that can be gleaned from unjust action without fear of penalty ought to be seized. Glaucon prevails upon Socrates to refute this popular view, and demonstrate that being just is something good in and of itself (Plato, *Republic II*, 358d). Only then would it not be “wretched and stupid” to live justly even when injustice would pay.

By demanding from Socrates a proof that a just life is good in and of itself, Glaucon seems to expect a particular kind of answer to the “Why be moral?” question. It is the answer the Plato gives by the end of Book IV and develops throughout the remainder of the *Republic*. According to Plato, we should all be moral because it is good for each and every one of us. A just person is happy, because to be just is to achieve a harmony of the soul (Plato, *Republic*, 443c – e). We sometimes have doubts about why we should be just only because we fail to appreciate what is good for us. This is what I call the classical answer to the “Why be moral?” question. Aristotle offers a similar answer in Book I of the *Nicomachean Ethics*, which I present in more in depth in
chapter 4 (section 4.3.1). The core idea is that eudaimonia—happiness or flourishing—is the highest good for any human being. Flourishing consists in rational activity in accordance with virtue. On Aristotle’s account, a virtuous life is a happy and excellent life for any human being. Once again, the “Why be moral?” question is answered by the thesis that justice and virtue are good for the individual him- or herself.

Another group of philosophers who have expressed influential answers to the “Why be moral?” question hail from the modern period. They are Thomas Hobbes, David Hume, and Immanuel Kant. In the coming chapters, I will discuss the relevant work of these thinkers and their contemporary followers. The Kantian answer differs from the classical answer in that it doesn’t claim that an ethical life promotes the agent’s own good or happiness. Instead, Kant argues that violating the moral law amounts to a form of incoherence or inconsistency. The basic idea of the Kantian position is that rational action requires impartiality (Brink, 1997). An agent acts from an impartial point of view as long as he follows the moral law as dictated by Kant’s Categorical Imperative. A rational agent must value the fact of his or her own rational agency. That is, rational agents value their own capacity to govern their choices and actions by reason. However, rational agency is instantiated not just in the self, but in others as well. So a rational agent must be impartial in the sense that he must be concerned to promote rational agency in whomever it exists. This is guaranteed if one acts as the Categorical Imperative directs, never treating rational agents as mere means, but only as an ends in themselves. Violating the moral law constitutes a disregard for the rational agency of others. In short, immoral
conduct is a failure to value that which one values, which is a form of rational incoherence or inconsistency. For Kant, then, we have reason to be moral, because moral conduct is a condition for being a rational agent.

The classical and Kantian answers to the “Why be moral?” problem are both committed to a thesis called *moral rationalism* (see chapter 3, section 3.1). According to moral rationalism, it is necessarily the case that if an agent morally ought to do A, then that agent has a reason to do A. In other words, moral requirements necessarily entail *reasons for action*. The next two philosophical figures who loom large in this project reject moral rationalism. They are Hobbes and Hume. Both of them believe that people are bound by moral obligations, but these obligations do not necessarily imply reasons for agents to honor them. Hobbes’s and Hume’s rejection of moral rationalism is a consequence of how they conceptualize a *practical reason* or *reason for action* (see chapter 1, section 1.0). Both of them are committed to a view I call *instrumentalism about practical reasons* (cf. chapter 1, section 1.1).² On this view, an agent has a practical reason to do action A if and only if A maximizes the satisfaction of his or her desires. It’s important to note, however, that instrumentalism is incompatible with moral rationalism. As we saw above, people’s desires sometimes favor ignoring their obligations for the sake of a preferred benefit. In those cases, one’s reasons for action do not side with morality. Thus, instrumentalism permits the possibility that one could have

² Although I treat Hume as an instrumentalist, some of his interpreters argue that he was in fact a *skeptic* about reasons for action. See Millgram (1995).
a moral obligation, without having a reason to honor it. This contradicts moral rationalism.

The instrumentalism of Hobbes and Hume return us to Glaucon’s challenge. If reasons for action depend on an agent’s desires, and an agent can have intense desires to act unjustly in pursuit of some advantage, then what reason is there to live ethically? Both Hobbes and Hume struggled to address this challenge. In *Leviathan*, Hobbes argues that moral rules constitute covenants adhered to by individuals in order to secure a stable, peaceful, cooperative and productive society. Such a society is advantageous for every agent. In particular, it advances what Hobbes considers to be the universal human desire for self-preservation. Thus, according to Hobbes, each member of society has reason to follow moral covenants on the condition that others follow them as well. If, on the other hand, other people violate moral covenants, then one ceases to have reasons to keep following covenants. As Gregory Kavka notes, “[t]o comply with moral rules unilaterally is to render oneself prey to others, and this, Hobbes urges, one is not required to do” (Kavka, 1985, in SL, p. 163). However, Hobbes considers the possibility that a figure he calls the fool could secure for himself the best of both worlds. He could exploit others by violating the moral covenants, while not risking retaliatory violations by others. That way the fool would enjoy all the benefits procured by moral covenants, while at the same time seizing every opportunity to advantage himself by violating the rules. Hobbes hears out the fool’s objection in this canonical passage (Hobbes, *Leviathan*, in SL, p. 614):

The fool hath said in his heart…that, every man’s conservation, and contentment, being committed to his own care, there could be no reason, why every man might
not do what he thought conduced thereunto: and therefore also to make, or not make; keep or not keep covenants, was not against reason, when it conduced to one’s benefit. He does not therein deny, that there be covenants…and that such breach of them may be called injustice, and the observance of them justice: but he questioneth whether injustice…may not sometimes stand with that reason, which dictateth to every man his own good.

In chapter 5 (section 5.4.1), I discuss Hobbes’s reply to the fool. The key idea behind this reply is that violating moral covenants is a bad wager. Although the fool may think he can exploit opportunities to violate covenants without provoking retaliatory violations by others, he’s likely to be mistaken (Darwall, 1995, pp. 76 – 77; Kavka, 1995). And so when the fool brazenly violates a moral covenant in pursuit of some advantage, he will find himself unexpectedly punished. Hobbes eventually concludes that the best policy for the fool is to abide by all moral covenants without exception. Therefore, Hobbes’s answer to the “Why be moral?” question is that for most people, most of the time, acting in accordance with moral covenants serves one’s desires.

Like Hobbes, Hume also defends an instrumentalist answer to the “Why be moral?” question, and in so doing he too runs up against Glaucon’s challenge.

Furthermore, Hume follows Hobbes in treating morality as a system of mutually advantageous rules called conventions (Darwall, 1995, p. 292). The prime example of a moral convention in Hobbes is justice, which dictates that one should leave others in the possession of their goods, provided that others act in the same manner (Ibid.). The main texts in which Hobbes develops his convention account of ethics are A Treatise of Human Nature and An Enquiry Concerning the Principles of Morals. In the Enquiry, Hume takes
note of Glaucon’s challenge through the mouthpiece of the *sensible knave* (Hume, *Enquiry*, 282):

[A] man, taking things in a certain light, may often seem to be a loser by his integrity. And though it is allowed that, without a regard to property, no society could subsist; yet according to the imperfect way in which human affairs are conducted, a sensible knave, in particular incidents, may think that an act of iniquity or infidelity will make a considerable addition to his fortune, without causing any considerable breach in the social union and confederacy.

The sensible knave claims to have the same talent as Hobbes’s fool in seeing “particular incidents” where violating moral conventions would not provoke retaliatory violations by others. In those incidents the knave can enjoy the advantages of “iniquity” while also benefiting from general compliance with moral conventions on the part of others. As I will discuss in chapter 5 (section 5.4.2), Hume’s reply to the knave complements and surpasses Hobbes’s rebuttal to the fool in various ways. In particular, Hobbes has no answer to Glaucon’s demand for a proof that a just life is good in and of itself, and not merely because of the extraneous honors and rewards it can afford. For Hobbes, the fool has reason to comply with moral covenants because he wants to avoid punishment. Hume, on the other hand, thinks that the knave neglects reasons one might have to abide by moral conventions for its own sake. As a matter of fact, Hume suggests, people are disposed to regulate their own conduct by the conventions of their society for no other reason than that the conventions are generally obeyed (Darwall, 1995, p. 292). In my terminology, people internalize the moral norms that prevail in their social environs. They thereby form desires to comply with moral norms simply because they want to comply.
In the tradition of Hobbes and Hume, this project defends an instrumentalist answer to the “Why be moral?” question. The present effort is intended to be a contribution toward a body of work produced by contemporary theorists who consider themselves heir to Hobbes and Hume. This literature is united by the thesis that *most individuals, most of the time, have reasons to lead an ethical life in virtue of the fact that moral conduct serves their desires and interests* (cf. Gauthier, 1986; Binmore, 1994, 1998, 2005; Kavka, 1985, 1995; Sugden, 2004). The existing literature has reached a remarkable degree of sophistication. Gauthier, Binmore, and Sugden, have applied game theoretic analyses to show how principles of distributive justice, property rights, norms of public contribution, and rules against harm might come to be followed by instrumentally rational agents whose only motivation is to satisfy their own desires. That by itself is a significant achievement. However, a limitation of the existing literature is that it fails to engage with philosophers who hold radically different conceptions of reasons for action. Whereas instrumentalists maintain that reasons for action are constituted by an agent’s desires, philosophers working in the tradition of Aristotle and Kant insist that reasons for action are entirely independent of any individual’s desires and/or motivations. The latter account of reasons is an important bulwark for moral rationalism. If true, it would explain how someone who has little or no desire to act justly might nevertheless have a reason to do so, solely because it is her duty. As the classical and Kantian answers to the “Why be moral?” question attest, moral rationalism has had an impressive pantheon of defenders. It cannot be casually dismissed. If the Hobbesian-Humean answer to “Why be moral?” is
to be persuasive, it must be shown that the instrumentalist conception of practical reasons which it presupposes is superior to alternatives. This will be my primary task in chapters 1 thru 4.

Existing Hobbesian-Humean rationales for the ethical life are also relatively thin on empirical evidence. They are, for the most part, concentrated on theoretical arguments, model-building, and speculative psychology. Hume’s thesis that people can acquire desires to conform to moral conventions for its own sake is in particular need of substantiation. However, I argue in chapter 5 that recent research in experimental game theory and in the psychology of social norms provides a wealth of evidence for a Humean moral psychology (see especially chapter 5, section 5.3). In my view, norms research can clarify and update the insights of Hobbes and Hume on the troubling question of how morality can be reconciled with individuals’ desires.

The following dissertation also tries to dispel a major anxiety about the Hobbesian/Humean answer to “Why be moral?” The worry is that both Hobbes and Hume reduce the notion of “being moral” to action in accordance with prevailing social conventions or norms. But the relationship between convention and morality can be quite contingent. Obviously, there are social rules which accord with morality insofar as they forbid theft and unprovoked harm. But social conventions can also be horrifically unjust. Consider apartheid, caste systems, and slavery. It may seem perverse to answer the “Why be moral?” question with the refrain that we all have reason to follow convention, when convention could just as well support reasons to injure and oppress. I offer a reply to this
challenge in chapter 5 (sections 5.5 – 5.6). My argument is that an individual can have reasons to challenge the prevailing conventions of one’s society and bring them into conformity with one’s moral ideals. As an example, I discuss successful campaigns to end the practice of female genital cutting (FGC) in some African communities. FGC is widely considered to be not only harmful to girls and women, but also a violation of their human rights (see in particular Nussbaum, 1999, in Feinberg & Shafer-Landau, 2008). Despite the fact that FGC is an age-old custom, women and men in several communities agreed to abolish it within just a few years. Finally, in section 5.6, I contend that norms favoring values of freedom, tolerance, and compassion are especially likely to emerge and endure in human societies. For these norms, I suggest, are more likely to promote people’s basic human needs than harmful or repressive norms. Thus, just as the gulf between individuals’ desires and their moral obligations is not as unbridgeable as Glaucon suggests, so there are forces that tend to bring morality and social convention into alignment.
1. Chapter 1: Practical rationality and the internalism-externalism debate

1.0 What is practical rationality?

Moral rules are practical, for they purport to govern our actions. At least part of what it means to “be moral” is to act in a moral way. Thus, by raising the “Why be moral?” question, we invite reflection about what one ought to do. Philosophers call the process of deliberation about action practical reasoning (Wallace, 2008). Practical reasoning consists in making normative judgments about the reasons one has to act. It is often contrasted with theoretical reasoning, which is thinking about the reasons one has to believe something. Thus the domain of theoretical reasoning is belief, whereas the domain of practical reasoning is action. With Thomas Scanlon, I define a reason as a consideration that favors or justifies something (Scanlon, 1998, p. 17). This is the normative interpretation of the term “reason.” Under the normative interpretation, it makes sense to say things like “My reason for believing that it will rain today is that the weather report indicated a high chance of showers.” Here, the weather report constitutes a reason which justifies the belief about impending rain. A different way to understand “reason” is the explanatory interpretation, according to which a “reason” is a cause of something. On the explanatory interpretation, I may sensibly say that “The reason I did poorly on the geometry test was that I believed the Pythagorean Theorem stated $a^2 + b^2 = c.$” In this context, the “reason” cited serves an explanatory function. The sentence tells us that my incorrect belief about the Pythagorean Theorem was the cause of my
performing poorly, but clearly it does not imply that I was justified in holding that belief.

Sometimes, one and the same consideration can serve both the explanatory and the justificatory functions: e.g., not only was my belief that it will rain today justified by the weather report, but my hearing the report was the cause of my forming the belief in the first place. We shall see, however, that the reasons which justify a particular belief or action may not be the ones that cause it, and vice versa. Unless otherwise noted, I shall hereafter use “reason” in the normative, rather than the explanatory sense of the term.

There are as many reasons as there are things which can be justified—reasons to act, reasons to believe, reasons to feel a certain way, reasons to value things, etc. Different reasons can favor inconsistent conclusions: I may have some reasons to believe in God, but other reasons to deny God’s existence. Also, one reason can be better than another in the sense that it counts “more strongly” in favor of something: I may have some reason to continue investing in my education, but other reasons that more strongly favor getting a job. Hence I have a better reason to get a job. If a reason counts more strongly in favor of something in particular than any alternative, it’s natural to say that this is one’s best reason.

I define a rational creature as one whose beliefs, actions, emotions, etc. are responsive to the best reasons. Conversely, a subject who doesn’t act, believe, feel, etc. according to what is favored by the best reasons is, in one way or another, not fully rational. The phrase “not fully rational” is intended as a general descriptor for a continuum of ways in which a creature can be unresponsive to reasons (cf. Scanlon,
At the extreme, some philosophers reserve the term “irrational” for cases where a person’s beliefs, actions, etc. fail to conform to his or her own judgments about the reasons he or she has. Scanlon uses the example a person who continues to believe something that she herself judges there to be strong reason to reject. Some people admit to having little or no reason to believe in God, but believe it nonetheless. A moderate degree of unresponsiveness to reasons can be captured by the predicate “open to rational criticism.” This would involve errors in judgment about the reasons one has—for instance, when a person is convinced that he or she has reason to believe in creationism, but only based on some fallacious inference from available evidence. It is not quite right to call such persons irrational because, by hypothesis, if they were shown the error in their reasoning, they would be inclined to revise their beliefs. At the same time, such persons may not be fully rational because, despite their best intentions, they are prone to fallacious reasoning. Finally, the term “non-rational” might be confined to the sorts of creatures that are incapable of responding to reasons at all. Bacteria are non-rational creatures, because they lack the capacity to think, act, feel, etc. in any way, not to mention ways that are justified by reasons. I shall employ the phrase “not fully rational” as a catch-all term in reference to a creature’s being “irrational” or “open to rational criticism” or even “non-rational.”

In answering the “Why be moral?” question, I will be concerned to identify practical reasons, or reasons for action, which favor acting morally. An agent who fails to act according to what is favored by the best practical reasons is, I assume, not fully
rational. Conversely, a *practically rational agent* is one who successfully identifies his or her reasons for action, and reliably acts in accordance with *best* reasons for action. Since my task is to specify the conditions under which the best practical reasons favor moral action, I will have to presuppose a *theory of practical rationality*. Just as logic provides a set of principles for the justification of belief, so a theory of practical rationality offers a set of principles for the justification of action. I will call these *principles of practical reasoning*.

In this chapter, my first task is to elucidate what I take to be the most tenable theory of practical rationality. This theory is known in the philosophical literature as *instrumentalism*. In sections 1.1 – 1.2, I summarize the basic contours of instrumentalism. In sections 1.3 – 1.4, I differentiate instrumentalism from a related view called *reasons internalism*. Instrumentalism, I shall argue, is a particular version of *actual* reasons internalism. I also distinguish two competing forms of reasons internalism—*actual* versus *counterfactual* internalism—and I offer an interpretation of Bernard Williams’s classic argument for actual reasons internalism. This task will be important, since in later chapters I will draw upon Williams’s argument in order to support an instrumentalist account of the rationality of *moral* action. In section 1.5 I characterize another view, *externalism*, which rivals all forms of internalism. Here I defend Williams against the common charge that his argument for reasons internalism is question-begging.

My overarching goal in this chapter will be to explain various points of disagreement among the perspectives on practical rationality, so that I may refer back to
them in later stages of this project. A very important point of disagreement concerns the relationship between practical reasons and the explanation of action. Internalists assume that practical reasons must not only justify action, but also explain it. Within the internalist camp, actual internalists and counterfactual internalists go on to debate how a process of practical reasoning can motivate action. Externalists, on the other hand, either try to weaken this assumption or (less commonly) they reject it altogether. If the internalists are right to think that reasons for action must be able to explain action, then the best theory of practical rationality has to include a true psychological account of how practical reasons can bring agents to act. However, if (some) externalists are right to suppose that psychology is irrelevant to practical rationality, the burden is on them to explain away Bernard Williams’s famous observation that practical reasons figure centrally in explanations of action. In chapter 2, I argue that the externalists cannot discharge this burden: to be a practically rational agent just is to be the sort of creature whose behavior can be explained by practical reasons. But first, an exposition of the various positions is needed to clarify what is at stake.

1.1 Instrumentalism about reasons

Philosophers have debated numerous competing theories of practical rationality. The one I embrace has come to be known as instrumentalism about reasons (hereafter simply instrumentalism). The central thesis of instrumentalism is that practical reasoning consists of “means-ends” reasoning about how best to achieve one’s desired ends (cf.
Millgram, 2001, p. 4). Though there are multiple specific versions of instrumentalism, I offer a generic characterization of the view summarized by the following claims:\(^1\):

**Instrumentalism about reasons**

*Definitions:* Let an *end* be a goal that an agent acts to achieve. For any action *A*, let *A* be a *means* to achieving end *E* if and only if *E* is a logical or causal consequence of *A*. A means to end *E* is an *optimal* means for an agent *S* if it is no less effective and efficient for *S* than any alternative means to *E*. Agent *S* has an *instrumental reason* *R* to perform act *A* if and only if (and because) there is some end *E* such that *S* desires *E*, and *S* justifiably believes that *A* is an optimal means to achieving *E*.

(I-1) All reasons for action are instrumental reasons.

*Definitions:* Suppose *S* has two desired ends, *E* and *E*'. Then *E* and *E*’ are *mutually exclusive* ends if and only if *S* cannot meet both *E* and *E*'. An agent’s *feasible set* of ends is the set of ends which the agent has sufficient resources (e.g. time, money, energy, etc.) to achieve. Let the agent’s feasible set of ends be

---

\(^1\) I draw this characterization of instrumentalism from Elster (1986, 2007), Millgram (2001), Fumerton (1990), and Brandt (1983).
represented by “{E}”. The risk attitudes of an agent are the effects that the risk of not achieving a given end has on the agent’s desire(s) for that end.

(I-2) Suppose that agent S has a set of instrumental reasons, {R}, to achieve ends in her feasible set {E}. Let R and R’ be two distinct members of {R}, such that R is a reason to A in order to achieve E, and R’ is a reason to A’ in order to achieve E’. Suppose that E and E’ are mutually exclusive elements of {E}. Then, given S’s risk attitudes and justified beliefs about the relative probabilities of achieving E versus E’,

(1) R is a better reason than R’ for S if and only if S has a stronger desire to meet E than E’, and

(2) R is the best reason for S if and only if R is a better reason than any other reason in {R}.

Definition: S is instrumentally rational if and only if, for any instrumental reason R to do A, if R is S’s best reason, then S is motivated by R to A.

(I-3) Agent S is practically rational if and only if S is instrumentally rational.
(I-4) If $S$’s best instrumental reason $R^*$ is to $A^*$ in order to achieve some end $E^*$, then $S$’s justified belief that $A^*$ is an optimal means to achieving $E^*$ is based on an optimal amount of evidence collected by $S$.

(I-5) No desired end $E$ of $S$ is open to criticism from the point of view of practical rationality, except insofar as $S$ may have a better instrumental reason to pursue some other end, $E'$.

Instrumentalism claims that the rationality of an action depends solely upon its relationship to the ends, beliefs, and desires of the agent. The ends of an agent are the goals he or she strives to achieve (Schmidtz, 2001, p. 237). According to instrumentalism, a reason favoring some action is engendered by the fact that the relevant action is an optimal means to satisfying the agent’s desired ends (cf. I-1 above). Also, the stronger the desire which is satisfied by the action under consideration, the better the reason one has to take the action (cf. I-2 above). The instrumentally rational agent is one who aims at the maximal satisfaction of his or her most strongly desired ends (cf. I-3 above).

Consider the following illustration. Imagine Adam, alone in the Garden of Eden. Adam has a taste for the local fruit, of which there are only two types: figs and mulberries. Corresponding to each alternative, he has a choice between exactly two

---

2 The case of “Adam in the Garden” is inspired by Hollis (1987).
actions: he could cross the stream to the fig tree, or he could trek up the hill to find the mulberry bushes. Each action, let’s assume, is sure to bring about its corresponding objective, and there is no alternative to each means of achieving either objective.

Accordingly, crossing the stream is by hypothesis an optimal means to finding figs, and trekking uphill is likewise an optimal means to collecting mulberries (cf. I-1 above).

Furthermore, let us assume that Adam knows, and therefore justifiably believes, these facts about which actions are optimal means to achieving which ends. Suppose, too, that Adam cannot gather figs and mulberries at the same time. It follows that if eating figs and eating mulberries are both ends that Adam desires, they are mutually exclusive ends. However, as much as Adam likes mulberries, he likes figs even more. Thus we may say Adam has both a desire to eat figs and a desire to eat mulberries, but he has a stronger desire for figs. Because Adam wants to eat figs, and because crossing the stream is an

---

3 That different desires can have a range of intensities is, I think, a familiar fact of conscious experience. This notion of the intensity of desires allows for desires to be compared both ordinally and cardinally. When desires are compared ordinally, they are ranked on a hierarchy from weakest to strongest. However, merely ordinal comparisons do not provide information about how much stronger one desire is relative to another. Thus we can sensibly say that Adam desires figs more strongly than he desires mulberries, and this can be understood without further information about how much more strongly Adam desires figs than mulberries. In that case, strength can be represented with the greater-than (“>”) or less-than (“<”) symbols. If Adam has a stronger desire for figs than for mulberries, then we can represent this relation symbolically with the expression “F > M.”

Of course, we also think of a single desire as having stronger or weaker intensity. This aspect of strength allows for cardinal comparisons among desires. For instance, if the strength of Adam’s desire for figs is mild, this would impact his behavior. Perhaps he could be easily persuaded to look for mulberries instead, if it would save him some time or if he could expect to get a larger fill of mulberries. But if Adam’s desire for figs is particularly strong, it would perhaps take the prospect a much larger bounty of mulberries to convince him to abandon his search for figs. To represent the intensity of a desire cardinally, numbers are useful. If we represent the intensity of Adam’s desire for figs by the number 4, and his desire for mulberries by 3, this suggests that Adam’s desire for figs is only slightly stronger than his desire for mulberries. But if Adam’s desire for figs can instead be represented by the number 30, this indicates that Adam has a much stronger desire for figs than mulberries (10 times stronger, in fact).
optimal means to getting figs, Adam has an instrumental reason to cross the stream. But notice that Adam also has a desire, albeit a weaker one, to eat mulberries. In addition, Adam knows that trekking up the hill is an optimal means to reaching the mulberries. Therefore, it can be said that Adam also has an instrumental reason to trek up the hill.

So there is Adam, faced with a choice between acting on two instrumental reasons for action. Instrumentalism holds that Adam will not act rationally just by acting on any of his instrumental reasons. Rather, his rationality depends on his having the capacity to act on the best of his instrumental reasons. The stipulation that any instrumentally rational agent must be motivated by his or her best instrumental reasons is expressed by thesis (I-3) above.

Thesis (I-2) offers a definition of a best reason for action. Instrumentalism is not unique among theories of practical rationality in positing that one can have reasons which favor a particular action, and at the same time have reasons favoring incompatible actions. And like other major theories of practical rationality, instrumentalism maintains that when reasons conflict, a reason in favor of act \( A \) may have a qualitative degree of “weight” or “force” vis a vis reasons which favor not doing \( A \). A practical reason that is “weightier” than another may be said to be the better reason, and the “weightiest” practical reason is simply the best practical reason the agent has. Applied to Adam’s choice between eating figs or mulberries, Adam’s best reason depends on several factors.

\[ \footnote{Intuitively, a similar relation holds among reasons for belief. One may have some reasons to believe the Republican candidate will win the next election, while at the same time having stronger, better, weightier, or more forceful reasons for believing that the Democrat will win.} \]
Thesis (I-2) says that if Adam more strongly desires figs than mulberries (i.e., if Adam prefers figs to mulberries), then he has a better reason to cross the stream than to trek up the hill. On the other hand, instrumentalists recognize that actions almost always involve an element of risk. An action may result in more than one possible outcome, and the individual may not be able to predict with certainty which outcome will come to pass. However, the agent may be able to evaluate the probability that each possible outcome will occur.\(^5\) Intuitively, the rationality of an action depends somewhat on the probability or improbability that it will succeed in bringing about the intended goal. Instrumentalism can accommodate this intuition. For instance, suppose it is some time after the Fall, and Adam has been expelled from the Garden. He is condemned to toil for food, and he can no longer be sure that his every want will be met as it once was in Paradise. But suppose Adam faces the same choice between exactly two alternative actions: cross the stream to eat figs, or trek up the hill to eat mulberries. However, now that he’s outside the Garden, Adam cannot be certain that crossing the stream will lead him to figs or that trekking up the hill will lead him to mulberries. Under these circumstances it’s possible that Adam will have a better reason to eat mulberries, despite his stronger preference for figs. This will be the case when Adam justifiably believes his chances of eating mulberries are better than eating figs. Suppose, for instance, that Adam takes note of the season and concludes it’s improbable that the fig tree is currently bearing fruit. He would hate to go

\(^5\) Sometimes, the agent will not even be able to attach probabilities to the outcomes associated with his or her action. Decision theorists call this a situation of uncertainty (Hargreaves Heap, 1992, p. 349).
through the trouble of crossing the stream, only to find dried fig carcasses scattered about. He is, on the other hand, relatively confident that plump mulberries will be on hand up the hill. And remember: Adam does like mulberries. In situations like this, instrumentalism allows for the possibility that Adam has a better reason to trek up the hill for the mulberries than he has to cross the river for the figs. According to instrumentalism, practical reasons are sensitive the likelihood of achieving an intended goal because the strength of an agent’s desires depends on the probability that they can be satisfied. Thus instrumentalists would have us expect that Adam’s desire for figs diminishes as a function of the estimated improbability that he can acquire figs. And when the probability of finding figs is sufficiently lower than that of finding mulberries, Adam will form a stronger desire for mulberries than figs. Furthermore, if we maintain the assumption that Adam has a choice between exactly two actions—viz., crossing the stream or trekking up the hill—and no others, then whichever end Adam has the strongest desire to pursue will be the one he has the best reason to pursue.

In the last paragraph, it was assumed that risk has a dampening effect on the strength of one’s desire for an end. This assumption is reflected in the saying “A bird in hand is worth two in the bush.” If Adam is like most people, he will prefer a search for 100 figs that he is guaranteed to find over a search for 100 figs that involves some chance of finding nothing. The degree to which one’s desires are also affected by risk is also a function of one’s risk attitudes. What should Adam do if God offered him the choice between a search for 100 guaranteed figs, and a search for a cache of 200 figs where there
was a 50% chance that he would find nothing? If Adam were to actually undergo the risky search many times over and over again, in the long run he would retrieve on average about 100 figs. But this does not mean that Adam would be equally satisfied by either search. Adam’s preference between the two searches will depend on his risk attitudes. If Adam is risk-averse, the risk of coming up empty-handed will deter him more than he’d be encouraged by the prospect of finding a 200-fig jackpot. In that case, Adam will prefer the sure bet of 100 figs to the risky search for 200 figs. On the other hand, if Adam is risk-loving, the prospect of the 200-fig jackpot will encourage him more than the risk of finding nothing will deter him. So, he’ll prefer the risky search for 200 figs to the sure bet of 100. And if Adam is risk-neutral, he will be indifferent between a search in which he would receive a guaranteed payoff and a risky search where, in the long run, he could expect to receive the same payoff.

Suppose now that Adam has more than two instrumental reasons to act in pursuit of more than two mutually exclusive ends. Then according to (I-2), Adam’s best reasons for action will be the subset of his reasons which, if acted upon, would make the aggregate strength of Adam’s satisfied desires greater than the aggregate strength of desires that would be satisfied by his acting on any alternative subset of reasons. Thus if Adam is faced with a choice between acting on pair of instrumental reasons to cross the stream for the figs and smell the wildflowers, or a pair of reasons to trek up the hill for the mulberries and bathe in the hot springs, his best reasons would be specified by whichever set of ends—eating-figs-plus-smelling-wildflowers or eating-mulberries-plus-
bathing-in-hot springs—he has a stronger desire to achieve. These considerations suggest that the following principle can be derived from (I-2):

**Definitions:** Suppose the agent $S$ has a set of instrumental reasons $\{R\}$ to pursue a set of actions $\{A\}$ in order to achieve a set of desired ends $\{E\}$, and another set of instrumental reasons $\{R'\}$ to pursue a set of actions $\{A'\}$ in order to achieve a set of desired ends $\{E'\}$. Let $\{E\}$ and $\{E'\}$ be mutually exclusive sets of ends. Then

(I-2') \{R\} is the best set of instrumental reasons for $S$ if and only if the aggregate strength of $S$’s desires satisfied by achieving the ends in $\{E\}$ is greater than the aggregate strength of desires satisfied by achieving the ends in $\{E'\}$.

The simplest case in which (I-2') applies would be one where $S$ has only two mutually exclusive ends, $E$ and $E'$, and two corresponding instrumental reasons, $R$ and $R'$. Note that thesis (I-3) defines an instrumentally rational agent as one who acts on his or her best instrumental reasons. Therefore, (I-2') and (I-3) entail a general principle of practical rationality which is the foundational rule of instrumentalism:

**The Instrumentalist Dictum:** (i) For any agent $S$, $S$’s best (instrumental) reasons for action favor pursuing those actions which $S$ justifiably believes will maximize the aggregate strength of $S$’s satisfied desires; (ii) If $\{R^*\}$ is $S$’s best set of reasons
for action, then $S$ is *instrumentally rational* if and only if $S$ is motivated by $\{R^*\}$ to maximize the aggregate strength of $S$’s satisfied desires.

The Instrumentalist Dictum is often expressed as the principle that a rational agent should act to maximize his or her net *utility*. I shall sometimes make use of this language (particularly in chapter 5). However, a word of caution is in order. The term “utility” is most commonly associated with “expected utility theory.” This theory, pioneered by John von Neumann and Oskar Morgenstern, continues to be a widely-used explanatory device in the social sciences. Roughly speaking, “utility” is shorthand for “desire-satisfaction,” although the concept has been defined much more precisely in the canon of expected utility theory. Though some theorists wish to maintain that expected utility theory should only be considered a body of positive principles to be used for the explanation of action, others believe these same principles can be treated as normative criteria for the justification of action (cf. Hausman, 2000; Sen, 1982). I will not have anything to say about this debate here. However, I do wish to point out that the normative treatment of expected utility theory is one, *but only one* version of instrumentalism. Other versions of instrumentalism that do not rest on the particular axioms of expected utility theory are possible (see especially Sugden, 1991). In this project I will not be committed to the specifics of expected utility theory. My claim is rather that *some* version of

-----------------------------

6 For an elementary introduction to expected utility theory, see Elster (2007), ch. 11.
instrumentalism which satisfies theses (I-1) through (I-5) is the correct theory of practical rationality.

Having distanced my generic characterization of instrumentalism from the details of expected utility theory, it will nevertheless be convenient to adopt some of the latter’s terminology. Under expected utility theory, the rational agent is one who “maximizes utility” for him or herself. Similarly, when paraphrasing the Instrumentalist Dictum, I shall say that the instrumentally rational agent “maximizes utility,” or equivalently, “maximizes the satisfaction of his or her desires.” This manner of speaking may tempt one into the mistake of equating the “satisfaction” of a desire with pleasure. Although one may take pleasure in satisfying a particular desire, pleasure and desire-satisfaction can come apart. For instance, I may take no pleasure whatsoever in telling a friend some disappointing news, even while I desire that I be the one to inform him. Thus, in my view, a desire for end \( E \) is satisfied just when \( E \) is the case. This leaves it open for a desire to be satisfied even if it brings no pleasure. On the other hand, in chapter 3 (section 3.5) I consider the possibility that a desire is a state which necessarily involves affect. Pleasure is a kind of affect, but not by any means the only kind. I may want to be the one who tells my friend the disappointing news because I would feel bad for him, possibly even guilty, if he found out any other way. Such feelings of empathy and guilt should not be equated with pleasure or pain.

---

7 For more on the distinction between pleasure and desire-fulfillment, see Roger Crisp, 2008.
A key feature of instrumentalism concerns the *explanatory direction* of the relationship between practical reasons, on the one hand, and desires together with means-ends beliefs about the optimal means to satisfying desires, on the other. This feature distinguishes instrumentalism from several alternative theories of practical rationality. The definition of an instrumental reason from thesis (I-1) says that agent S has a practical reason $R$ to $A$ if and only if (*and because*) $S$ desires to attain $E$, and $S$ justifiably believes $A$ is an optimal means to attaining $E$. The insertion of “and because” indicates that the existence of $R$ is explained by $S$’s package of a means-ends belief and a desire to $E$. Instrumentalists generally deny the possibility that the direction of explanation might go the opposite way—i.e., that $S$ can have a practical reason $R$ which explains the existence of either $S$’s desire for $E$ or $S$’s means-ends belief that $A$ is an optimal means to $E$. For the instrumentalist, an agent has a practical reason in virtue of having certain kinds of psychological states, not the other way around. There are, however, alternative accounts of practical rationality that allow for the agent’s recognition of his or her own practical reasons to produce psychological states, such as desires (see in particular Nagel, 1970).

Instrumentalism also posits important connections between rational action and rational belief. The instrumentally rational agent must form beliefs about whether a given action is an optimal means to achieving a desired end. But such an agent will succeed in achieving her ends only if her means-ends beliefs are reliably true. Instrumentalists hold that means-ends beliefs are more reliable—more likely to be true—if they are justified by certain rationality criteria. This thought motivates the clauses in (I-1) and (I-2) which
stipulate that the agent must “justifiably believe” a course of action is an optimal means to attaining an end. According to one broadly accepted criterion, rational means-ends beliefs must be internally consistent, that is, free from contradictions. Adam cannot rationally both hold and reject the belief that there is a probability $p$ of finding figs if he crosses the stream. Additionally, rational beliefs must be justified by a sound deductive argument or inductive argument with a high degree of inductive probability. Further, rational beliefs must be caused by the available evidence, and so caused in a way that reliably makes a high proportion of them true (cf. Elster, 1986, pp. 13 – 14; 2007, p. 202). Adam cannot rationally believe that ripe figs are available across the stream because he dreamt of eating figs the night before. Instead, he ought to search for evidence that will allow him to reliably predict the location of his coveted figs (the season, the weather, etc.).

Beyond these few very general constraints, additional and more specific criteria for the rationality of means-ends beliefs are more contentious. But instrumentalism as a broad view must place some rationality requirements on beliefs about means to ends. The reason is that having true beliefs about which actions are apt to promote one’s ends contributes to instrumentally rational action (cf. Brandt, 1983, p. 155). Suppose Adam is perfectly willing and able to pick some figs, his favorite fruit, and the fig tree is replete with fruit. But if Adam falsely believes that the fig tree has died, and consequently opts to stay where he is, he will end up less satisfied than he would have been, had he known the fig tree was carrying fruit. Having a false belief hinders Adam from taking the action
which would optimally promote his most desired ends. But by the instrumentalist dictum, an instrumentally rational agent is one who acts to maximize the satisfaction of his desired ends. So in order for Adam to be an instrumentally rational agent, his beliefs must meet at least some epistemic criteria which conduce to truth.

What has been said concerning the rationality of means-ends beliefs allows for the possibility that an agent could have justified but false beliefs that some action is an optimal means to an end. Instrumentalists are then faced with the question of whether an agent may have best reason to follow a certain course of action, despite her having a false means-ends belief that the relevant action will optimally promote her ends. Many prominent instrumentalists are committed to answering “yes.” Richard A. Fumerton offers one thought experiment that gives intuitive support to this view (Fumerton 1990, p. 96). Suppose I have a strong desire for physical health and longevity. The preponderance of medical research shows me that maintaining a low-salt diet would keep my blood pressure low. However, suppose that medical researchers had grossly miscalculated the relationship between salt intake and blood pressure, and it turned out that a high-salt diet actually promotes lower blood pressure. In that case, maintaining a low-salt diet would frustrate my desires for health and long-life. Nevertheless, given the best evidence available, I would be perfectly justified in deferring to the expertise of medical scientists and believing that my desires would be served by a low-salt diet. In view of these considerations, the characterization of an “instrumental reason” presented above (particularly in I-1 and I-2) makes instrumental reasons contingent upon the justifiability
of the agent’s beliefs, rather than their truth or falsity. Throughout this project (particularly, chapters 2 and 5), I shall follow Fumerton’s evidence-relative account of practical reasons.

Another condition of instrumentally rational action, indicated in thesis (I-4), is that the rational agent should invest an optimal amount of resources in acquiring evidence for her means-ends beliefs. It’s already been emphasized that instrumentally rational agents must base their means-ends beliefs on evidence. But the relevant evidence can be elusive, and it may take some effort on the part of the agent to search for it. It is a fact of mortal existence that actors have only finite resources—e.g., money, time, energy, etc.—to devote to the acquisition of information and knowledge. Consequently, a balance must be struck: if the actor allocates too much of her resources to collecting information about how to attain just one end, then she will inhibit her own ability to meet other ends she strongly desires; on the other hand, if the agent allocates too little toward getting information about one strongly desired end, she risks not informing herself well enough to attain that end.

Jon Elster has commented extensively on this as aspect of instrumental rationality, particularly in Elster (1986) and Elster (2007). He takes the optimal investment in the search for information to depend on the desires of the agent, the agent’s background beliefs about the situation, and the results of the search itself (Elster, 2007, pp. 205-06). First, the stronger one’s desire for a certain end, the more one ought to invest in searching for information about the best means to achieve that end. If I care more about staying in
good health than I do about owning fine wines, then I should invest more resources in researching exercise and diet regimens than I do in literature about wine. Second, the agent’s background beliefs specify the expected costs and benefits of obtaining new information. If acquiring a new piece of information about the best means to achieving end $E$ precludes opportunities to pursue other ends of mine, and if these opportunity costs are sufficiently high, I should not invest more resources in information about $E$. I may want very much to be the proud owner of fine wines, but beyond a certain point it would not be rational for me to devote more resources to researching the art of wine collecting if this would keep me from doing other things I love—like studying philosophy. Third, the optimal amount of resources to invest in information search may also be determined by the results of the search itself. Elster cites the example of testing new medications, where the experimental medicine is given to a treatment group but withheld from a control group. When the product turns out to be successful after repeated trials, clearly it would make sense to make it publicly available. Beyond a certain point, continuing to withhold the drug from the general population for the sake of further testing would thwart the primary goal of curing the sick.

According to Elster, “[e]very decision to act can be seen as accompanied by a shadow decision—the decision about when to stop collecting information. The rationality of the former depends on the latter” (Elster, 1986, p. 14). Elster calls the search for information a shadow action, because it presents the agent with a secondary choice problem which must be resolved prior to the primary problem. The solution to the
secondary problem can determine the solution to the primary problem. Before I can act in pursuit of end $E$, I must (if I’m instrumentally rational) select the optimal means to achieving $E$. But the opportunity costs of searching for the optimal means to $E$ may be so great that I would prefer to pursue other ends that I desire. Under those circumstances, I would have what we might call a best secondary reason to collect information relevant to finding the optimal means to $E$ only up to a certain point. Ideally, this is the point at which the marginal benefit of collecting further information about how to achieve $E$ equals the marginal cost—or to use terminology employed here, when the additional strength of the desires satisfied by acquiring additional information equals the additional strength of the desires satisfied by not acquiring additional information (cf. Elster, 1986, p. 14). Once I’ve obtained this optimal amount of information, I should assess my best reasons for action based on that quantity of evidence. However, in chapter 5 (section 5.4.2), I note that the “marginal cost equals marginal benefit” rule may not be the best one for determining the “optimal” amount of information about how to achieve one’s ends. In any case, the point for now is that (I-4) should be applied to the interpretation of the clause in the Instrumentalist Dictum stating that the instrumentally rational agent must “justifiably believe” that an action will maximize desire-satisfaction.

The final key tenet of instrumentalism is perhaps the most contentious. It is stated in thesis (I-5) above. (I-5) is the basis for the clause in (I-1) which states that all normative reasons for action are instrumental reasons, and for the clause in (I-3) which states that all and only practically rational agents are instrumentally rational agents.
Instrumentalists hold that the normative function of practical reasoning is restricted to the evaluation of means-ends beliefs about how agents ought to go about satisfying their strongest desires. It follows that the Instrumentalist Dictum is the only principle of practical rationality there is. On an instrumentalist conception of practical reason, an agent is open to rational criticism if she fails to pursue her most strongly desired ends, or if she fails to select the optimal means achieving her ends, or if she fails to hold justified beliefs about the optimal means to achieving her ends. But instrumentalism does not, and indeed it cannot, provide a normative basis for evaluating the ends one rationally ought to desire unless there are other ends whose achievement would produce a greater weight of satisfied desires (cf. Fumerton, 1990, pp. 94-95). Here it helps to introduce a frequently-made distinction between two types of ends: final ends and instrumental ends. An instrumental end is an end which an agent desires because its attainment is a means to fulfilling some other end. A final end is an end which an agent desires for its own sake (i.e., intrinsically) (cf. Schmidtz, 2001, p. 238; Fumerton, 1990, p. 94). Someone may have desires for rigorous exercise and low-fat foods, not because he or she necessarily enjoys these things more than fatty foods and lazing about the house, but because they are means to maintaining good physical health and a long life. In this case, the desires for exercise and low-fat foods are desires for instrumental ends. But if the person in question desires health and longevity simply for their own sakes, and not because the fulfillment of these desires would promote another end, then the desires in question are desires for final ends.
The five instrumentalist theses broached at the beginning of this section allow for normative criticism of instrumental ends, but not of final ends. Traditionally, languages have been taught in the classroom by making students memorize grammar rules. Studies have shown that this method is ineffective, especially when compared to having students speak and write in the language. Suppose a teacher desires to teach her students the French language as a final end. She is a serious professional, and on matters related to her profession her desire to teach French genuinely trumps all other concerns she might have. Unfortunately, this teacher is untutored in educational psychology. She erroneously but sincerely believes that compelling her students to master French grammar is the best way to get them to learn French. Alternatively, this teacher may be well-informed about the psychology data, but she has such a strong taste for tradition that she clings to the old grammar-oriented method. In either scenario, the teacher’s desired goal of getting her students to master French grammar can be described as an instrumental end aimed at her final end. This instrumental end can be criticized from the perspective of instrumental rationality on several fronts. The French teacher may be said to lack a justifiable means-ends belief that grammar drills are an optimal means to teaching a language. Either she fails to consider the available evidence in forming her belief, or she fails to invest enough effort in attaining the relevant evidence. In the alternate scenario, where the teacher is well-informed but still refuses to change her methods due to her conservative sensibilities, she is not fully rational because she is unmoved by the fact that she has a better (instrumental) reason to do otherwise.
By contrast, instrumental reasons for action are defined with reference to the agent’s final ends. Consequently, the bare fact that an agent has certain final ends, or certain such ends rather than others, cannot be criticized except by the charge that the agent has better instrumental reasons to pursue different final ends. If playing chess is among a person’s final ends, the instrumentalist cannot criticize this person for merely wanting to play chess, provided that no more strongly desired end is foregone by having or acting on this chess-playing want. On the other hand, if some more strongly desired end is excluded by acting on the desire to play chess, then the instrumentalist can say the person in question has a better instrumental reason either not to act on the desire or to abandon it altogether. Ultimately, this is the only basis on which an instrumentalist can evaluate desires for final ends. Thesis (I-5) has been forcefully criticized by opponents of instrumentalism, because it allegedly leads to tensions between instrumentalism and powerful normative intuitions about both practical rationality and morality. First of all, it appears to exempt from criticism certain ends that seem downright irrational for anyone to have—for instance, the overwhelming desire to eat saucers of mud (Anscombe, 1957). Secondly, it opens the possibility of tension between practical rationality and morality—two normative frameworks which many thinkers believe must be continuous with one another. Many people are disinclined to certify as rational the pursuit of morally abhorrent ends like rape or genocide, and yet if we assume that people can and do have strong desires for such depravities as final ends, instrumentalism commits us to the view that it can be rational to be wicked. These issues constitute the focal points of dispute.
between instrumentalists and their opponents. The main objective of this project is to vindicate the instrumentalist view on these matters. In chapter 4, I present David Schmidtz’s more sophisticated model of *reflective instrumentalism*, which does permit the evaluation of final ends. In chapter 5, I flesh out reflective instrumentalism in an effort to show that most human beings have compelling instrumental reasons to be moral.

### 1.2 Historical origins of instrumentalism

Now that the basic theses of instrumentalism have been explained in some detail, I’ll go on to characterize rival conceptions of practical rationality. But first: a historical aside. Instrumentalism has a distinguished pedigree in the history of philosophy. The two most important figures to have been interpreted as instrumentalists are Thomas Hobbes and David Hume. In his account of Hobbes’s views on practical rationality, Stephen Darwall suggests that Hobbes held that the *normative force* of practical reasons emanates from the laws of nature (cf. Darwall, 1995, pp. 57 - 59). Hobbes accepted Grotius’s thesis that a law of nature is a “dictate of right reason.” For Hobbes, reason dictates those actions which are the means necessary to fulfilling an agent’s desired ends. The normative force of any reason for action consists entirely in an agent’s having a desired end which is served by the relevant action. In particular, Hobbes thought it to be a

---

8 I understand the “normative force” of a consideration as the degree to which the consideration *ought* to affect our thoughts, actions, feelings, etc. Of course, normative force does not actually affect all people in the same way—Hitler certainly didn’t appreciate the normative force of the duty to preserve human life. Nevertheless, the force of normative considerations is ordinarily thought to be an objective property which exists even if some people happen to be insensitive to it. Hitler may have cared little about preserving human life, but we still think he violated an important (i.e., forceful) moral obligation.
fact of human nature that all people desire self-preservation. Therefore, all human beings have reason to pursue actions which are necessary means to ensuring their own survival. That said, Hobbes was never committed to the existence of reasons as *sui generis*, ontologically basic normative properties. He did not take the normative force of practical reasons to be a property in and of itself, since his metaphysics allowed only for the existence of matter and its modifications (Darwall, 1995, p. 58). Rather, Hobbes equated the normative force of practical reasons with the subjective, psychological experience of the “motive force” which transfers, through practical reasoning, from one’s desired ends to the means necessary to fulfill those ends (see especially Darwall, 1995, pp. 58 – 59). From an impersonal perspective, this transfer of motivational force via practical reasoning was for Hobbes “no more inherently normative than a fact about what must happen if water is to boil” (Darwall, 1995, p. 60). On the other hand, from the perspective of a rational human agent, the fact that an action is a means to the agent’s desired ends will be experienced as having normative force. Such a person will not only feel compelled to act on a desire for self-preservation, but will also judge the actions she deems necessary for her survival to be ones that she *ought* to undertake. Evidently, Hobbes held that human beings have a normative *concept* of a reason for action which applies to whatever means are necessary to achieving one’s desired ends.  

9 But his materialist metaphysics prevented him from holding that this concept has any reference to

---

9 Hobbes seems to hold a similar view concerning other normative concepts like the concept of *goodness*. In *Leviathan*, chapter 6, he suggests that something is called “good” when it is the object of desire, and something is called “evil” when it is the object of an aversion (cf. Kavka, 1986, p. 292).
other than a psychological process. In other words, on a Hobbesian picture, practical reasoning consists in deliberation about *instrumental* reasons for action.

Hume’s status as an instrumentalist is more contentious. Some passages in *A Treatise on Human Nature* and *An Enquiry Concerning the Principles of Morals* suggest that practical reasoning consists only in identifying means to desired ends (cf. Audi, 2006, chapter 2). However, other passages give the impression that Hume was in fact a *skeptic* about practical reasoning. Skepticism about practical reasoning is the view that there are no legitimate forms of practical reasoning—no normatively authoritative principles dictating how one ought to act—and consequently no reasons for action whatsoever (cf. Millgram, 1995; Millgram, 2001, p. 3). Those who read Hume as an instrumentalist tend to emphasize passages such as the following (Hume, 3.1.1.12, in Norton & Norton, 2000, p. 295):

> It has been observ’d, that reason, in a strict and philosophical sense, can have an influence on our conduct only after two ways: either when it excites a passion by informing us of the existence of something which is a proper object of it; or when it discovers the connexion of causes and effects, so as to afford us means of exerting any passion. These are the only kinds of judgments which can accompany our actions…and it must be allow’d, that these judgments may often be false and erroneous.

The view suggested here is that practical reason consists of making inferences among beliefs about the relation between desired ends and the actions necessary to attain ends. Conceived this way, practical reasoning can affect action in one of two ways: by informing agents of the nature or existence of something which is desired (the object of a passion), or by supplying agents with information about available means to desired ends.
Furthermore, the passage suggests that practical reasoning can be erroneous in several ways: it can misinform us about the nature or existence of an object desired, and it may be mistaken about how a desired end is to be achieved (cf. Audi, 2006, p. 47). In passages like these it sounds as though Hume did have a substantive conception of practical reasoning, and that he construed practical reasoning along instrumentalist lines. Indeed, Hume has been so closely associated with instrumentalism that many commentators have referred to the view as “the Humean conception of rationality” or “the Humean theory of reasons” (cf. Sugden, 1991; Finlay & Schroeder 2009; Parfit & Broome, 1997).

### 1.3 Instrumentalism and internalism about reasons

In recent philosophy, some have characterized instrumentalism as a species of a more general family of views which bear the name *reasons internalism* (cf. Finlay & Schroeder, 2009; Shafer-Landau & Cuneo, 2007, p. 283). There are numerous versions of reasons internalism, of which instrumentalism is just one, but I regard all of them as particular specifications of this schematic formula (cf. Shafer-Landau, in SLC, pp. 312 – 313; Robertson, 2001, p. 130; McNaughton & Rawling, 2004, p. 113):

*Reasons internalism*

**Definition:** An agent $S$ has an *effective* motivation $M^*$ to do action $A^*$ if and only if $M^*$ is a sufficient condition for $S$ to do (or attempt) $A^*$. 

(RI) For any agent $S$, and some rational principle $P$,

(i) if, according to $P$, there is a reason $R$ for $S$ to do action $A$, then if $S$ is practically rational, $S$ has some motivation $M$ to $A$, and

(ii) if, according to $P$, $R^*$ is a reason to do action $A^*$, and $R^*$ is $S$’s best reason for action, then if $S$ is practically rational, $S$ has an effective motivation $M^*$ to $A^*$.

Reasons internalism, or (RI), holds that a necessary condition for an agent to have a reason to $A$ is that the agent has some motivation or another to do $A$, provided that the agent is practically rational. The conditional in the “then” clause of (RI) implies that a

\[\text{It’s important to distinguish reasons internalism from other, quite different views that have also been called “internalism.” Reasons internalism is a thesis about the relation between an agent’s practical reasons and motivations. It states that an agent’s having a motivation is a necessary condition for the existence of a practical reason (cf. Darwall, 1997, in DGR, p. 308). But another view which has been called “internalism about reasons” asserts a necessary connection between one’s practical reasons and one’s moral duties, viz.} \]

\[(MR) \text{ Necessarily, if agent } S \text{ morally ought to do } A, \text{ then there is a reason for } S \text{ to do } A\]

This thesis is a central point of contention in debates concerning the relationship between practical reason and morality. We shall revisit it soon enough. But to prevent any confusion, we’ll follow the alternative convention of calling this thesis moral rationalism (“MR” for short) (compare Brink, 1989, p. 40, Darwall, 1997, in DGR, p. 306, with Finlay & Schroeder 2009 & Schroeder, 2008).

Yet another family of theses to be distinguished from reasons internalism makes claims about the relation between an agent’s own judgments or beliefs about her moral duties, and the motivations she has or rationally ought to have. Following Stephen Darwall, one such thesis may be dubbed moral judgment internalism (Darwall, 1997, in DGR, p. 308):

\[(MJI) \text{ Necessarily, if } S \text{ judges (or believes) that she morally ought to } A, \text{ then this judgment produces some motivation for } S \text{ to } A.\]

As Darwall notes, moral judgment internalism often appears in non-cognitivist arguments for the conclusion that moral judgment is a motivational state instead of a “cognitive,” i.e. belief-like state. (MJI) tells us, for example, that if I judge that I morally ought to give to charity, then because of this judgment I will have some motivation to give to charity (Darwall, 1997, p. 308; Shafer-Landau & Cuneo, 2007, p. 226).
practically rational agent is one who is motivated to act according to some rational principle or principles, which I label $P$. The practical reasons of an agent are individuated and ranked from best to worst by $P$. Reasons internalism does not require that a rational agent with a reason to $A$ will always be motivated to actually do $A$. Rather, as (RI-i) suggests, she must have some motivation to $A$. But she may have other, stronger motivations which cause her to act otherwise. On the other hand, as (RI-ii) stipulates, if a rational agent’s best reason is to $A$, then not only will she have a motivation to $A$, but this motivation will be effective in causing her to $A$ or at least attempt to $A$. (RI) is consistent with the possibility that the agent can have multiple reasons which $P$ ranks as the “best.” In that case, the agent will be rational as long as she is effectively motivated to act on all her best reasons. (RI) also allows for the possibility that an agent can fail to be motivated to do what she has best reason to do, with the caveat that in such a case the agent would

In addition, Michael Smith adverts to a proposition that draws a connection between an agent’s motivations and her judgments about her own reasons for action (cf. Smith, 1994, p. 148):

(RJI) Necessarily, if agent $S$ judges (or believes) that she has a reason to $A$, then either this judgment produces some motivation for $S$ to do $A$, or $S$ is not practically rational.

The relevant judgments here do not concern any moral obligations the agent has. Instead, the proposition refers only to the agent’s own judgments about her reasons for action. Hence, we may label this thesis judgment internalism about reasons, or “RJI.”

These propositions must not be conflated with reasons internalism. It’s perfectly consistent to accept (RI) while accepting, rejecting, or taking no position on (MR), (MJI), or (RJI).

11 It’s commonly assumed that an agent acquires a motivation to $A$ because of recognizing, through rational practical deliberation, that s/he has a reason to $A$. I’m inclined to reject this, because I hold that it’s possible for an agent to be motivated to act in accordance with rational practical deliberation, without being conscious of this deliberation or of its results. Therefore, I prefer to characterize rational agency as a propensity to be motivated by one’s reasons for action, whether or not one is conscious of them.
not be fully practically rational. Thus, an equivalent formulation of (RI) says that for any agent $S$, if (according to $P$) there is a reason $R$ for $S$ to do action $A$, then either (a) $S$ has some motivation to $A$ or (b) $S$ is not fully practically rational.

In the (RI) schema, $P$ is the principle (or group of principles) which determines what practical reasons agents have, and which among those reasons is the agent’s “best.” $M$ designates a motivation which a practically rational agent must be in as a necessary condition for having a practical reason to $A$. Different versions of reasons internalism result from varying specifications of $P$ and $M$. In particular, instrumentalism counts as a version of reasons internalism. Under instrumentalism, for an agent $S$ to have a reason to $A$, the relevant sort of motivation that $S$ must have is a desire to $A$. Furthermore, according to instrumentalism all it means for an agent to be practically rational is for her to be instrumentally rational (cf. I-3 and I-5 above). The rational principle, $P$, which determines the reasons of an instrumentally rational agent is the Instrumentalist Dictum. However, the Instrumentalist Dictum is not the only possible specification of $P$. Some reasons internalists argue for additional principles of practical rationality. Kantian internalists hold that the Categorical Imperative, in one or more of its formulations, counts as $P$ (see especially Korsgaard, 1986, 1996). Apart from these considerations, instrumentalism also makes a stronger claim than (RI). Again, (RI) asserts that having a motivation to $A$ is a necessary condition for having a reason to $A$, provided that one is

---

12 To be sure, there are other kinds of motivational states including, but not limited to, unconscious drives, instincts, moods, etc. Nevertheless, within the instrumentalist account desires play the central role in determining an agent’s reasons.
practically rational. But instrumentalism includes having certain motivations in a sufficient condition for having a reason. It holds that an action’s being such as to maximize the satisfaction of an agent’s desires is both a necessary and sufficient condition for the agent’s having a best reason to perform the action.

1.4 Actual reasons internalism

1.4.1 Procedural vs. substantive rationality

Instrumentalism is a version of reasons internalism. However, other versions of reasons internalism are incompatible with instrumentalism. A helpful way to clarify the differences among the various frameworks is to draw a distinction (based on Finlay & Schroeder’s 2008) between actual and counterfactual versions of reasons internalism. Though (RI) applies to both variants, the following formulations highlight their special features:

(ARI) *Actual reasons internalism:* If agent S has a reason R to do act A, then if S is procedurally rational, S has some motivation M to A.

(CRI) *Counterfactual reasons internalism:* If agent S has a reason R to do act A, then if S is substantively rational, S has some motivation M to A.
The difference between actual and counterfactual reasons internalism turns on the distinction between what’s been called *procedural* and *substantive* practical reasoning (Hooker & Streumer, 2004; Parfit & Broome, 1997):

*Procedural practical reasoning*: Suppose that agent $S$ has a practical reason $R$ to do act $A$. Then an operation of *procedural practical reasoning* is a series of inferences from $S$’s actual (or existing) motivations to the conclusion that $S$ has $R$, where this series of inferences is justified by some principle of practical rationality, $P(p)$. Let $P(p)$ be called a *procedural* principle.

*Procedural rationality*: Given that $S$’s having reason $R$ to $A$ is established by some procedural principle of practical rationality, $S$ is a *procedurally rational agent* only if $S$ has some motivation $M$ to $A$.

*Substantive practical reasoning*: Suppose that agent $S$ has a practical reason $R$ to do act $A$. Then an operation of *substantive practical reasoning* is a series of inferences from some *objective value* $V$ to the conclusion that $S$ has $R$, where this series of inferences is justified by some principle of practical rationality, $P(s)$. Let $P(s)$ be called a *substantive* principle.
Substantive rationality: Given that S’s having reason R to A is established by some substantive principle of practical rationality, S is a substantively rational agent only if S has some motivation M to A.

Procedural and substantive reasoning are distinguished by what might be called the “starting points,” or “basic inputs,” of practical deliberation. Proceduralists hold that all principles of practical reasoning are procedural. Procedural operations of practical reasoning are justified by procedural principles of practical rationality. A procedural principle of rationality can be thought of as a function from an agent’s actual or existing motivations to the agent’s reasons for action. For instance, the Instrumentalist Dictum is a procedural principle of rationality. It takes the agent’s actual desires, and identifies her practical reasons as those favoring actions which maximize the satisfaction of her desires.

Substantivists, on the other hand, hold that some (and perhaps all) operations of practical reasoning are substantive in nature. Substantive operations of practical reasoning are justified by substantive principles of practical rationality. A substantive principle of rationality is a function from objective values to an agent’s reasons for action. Objective values include things such as what is good for the agent to do, what is the agent’s duty, what is praiseworthy for the agent to do, etc. Different substantivists posit different specific relationships between objective values and reasons for action, but the general idea is that in certain situations, certain objective values ground certain reasons for actions. Such values are “objective” in the sense that they are thought to have
normative force independently of any particular agent’s judgments, motivations, or concerns.

Imagine, for instance, a young man raised by loving parents who provided him every opportunity to succeed. The man achieves much in life, in no small part because of his parents’ nurturing and support. But when his parents grow old and their health begins to decline, the man feels no inclination whatsoever to care for them. Despite all his parents have done for him, he fails to see why he should do them any favors. So he cuts off all contact with his parents, and goes on merrily with his own life while they languish in poverty and illness. It seems clear enough that this man has violated a moral duty to care for his parents. What’s more, this seems to be his duty regardless of the fact that he had no motivation to honor it. A substantivist about practical rationality may argue that this man had a reason to care for his parents solely because it was his duty, and in spite of his lacking a motivation to do it. The idea, here, is that people have practical reasons to do their duty simply in virtue of the fact that it’s right or worthwhile to do. Meanwhile, it’s assumed that people have moral duties whether or not they actually feel inclined to meet them. In sum, the “basic inputs” of procedural principles of rationality are the actual motivations of the agent, whereas the “basic inputs” of substantivist principles of practical rationality are objective values for the sake of which action is worthwhile. So-called objective values exist independently of the motivations of any particular agent.

Actual reasons internalism holds that there are only procedural principles of practical rationality. The view is named actual reasons internalism for two reasons: (1)
because, in specifying an agent’s practical reasons, procedural principles of practical reasoning take the actual motivations of the agent as basic inputs, and (2) because procedural principles place relatively weak constraints on the motivations an agent needs to have in order to be practical rational. For the actual internalist, a procedurally rational agent need not have any determinate motivations other than the ones he or she actually has, except for dispositions to be motivated by procedural operations of practical reasoning.

Unlike actual internalism, counterfactual reasons internalism assumes that there are substantive principles of practical rationality. The view is named counterfactual reasons internalism because substantive principles of practical rationality place relatively stringent constraints on the motivations that an agent would have if he or she were substantively rational. The substantively rational agent doesn’t only need to have dispositions to be motivated by operations of substantive practical reasoning; she also needs certain determinate motivations that she would develop as a result of being substantively rational. This is why the view is called counterfactual reasons internalism.

1.4.2 Instrumentalism as an actual internalism

Instrumentalism is a special case of actual reasons internalism, since its central principle of practical rationality—the Instrumentalist Dictum—is a procedural principle that determines an agent’s practical reasons based on her actual desires. Indeed, instrumentalism has it that the Instrumentalist Dictum is the only principle of practical rationality (c.f. I-5 above). Returning to the simple case of Adam in the Garden (section
1.1), an instrumentalist will say that Adam’s best practical reason is to cross the stream to reach the fig tree because he *actually* desires to eat figs more than anything else. The conclusion that Adam’s best reason for action is to cross the stream would be justified by the following practical syllogism:

(1) Adam’s strongest desire is for figs.
(2) Adam justifiably believes that crossing the stream is an optimal means to acquiring figs.
(3) Because of (1) and (2), Adam would maximize the satisfaction of his desires by crossing the stream.

*Procedural principle of practical rationality* (i.e., The Instrumentalist Dictum):
For any action $A$, $S$’s best practical reason is to $A$ if and only if $S$ justifiably believes that $A$ maximizes the satisfaction of $S$’s desires.

(4) Therefore, Adam’s best practical reason is to cross the stream.

Suppose that premises (1) - (3) of this practical syllogism are true. Then this argument will be sound—the truth of the premises guarantees the truth of the conclusion. The inference to the conclusion is justified by the Instrumentalist Dictum.

---

13 For expository purposes I make the cognitivist assumption that statements about normative practical reasons can be true or false. For a non-cognitivist account of practical reasoning, see Gibbard (1990, 2003). Gibbard’s account is noteworthy because it aims to show how inferential relations can hold among
Note that two of the premises in this syllogism (viz. (1) and (3)) mention Adam’s actual motivations: namely, a claim about Adam’s desire for figs, and a claim about the aggregate strength of Adam’s desires which would be satisfied by his crossing the stream. This is what it means for an agent’s existing motivations to constitute the “starting points,” or “basic inputs,” of an operation of procedural practical reasoning. Facts about the agent’s actual motivations must enter into a chain of practical reasoning which yields conclusions about what reasons the agent has. Thus, the Instrumentalist Dictum is a procedural principle of rationality. Insofar as Adam acts according to the Instrumentalist Dictum, he is instrumentally rational, and if we grant that the Instrumentalist Dictum is the only procedural principle, then he is also procedurally rational.

Here it may be objected that instrumentalism presents counterfactual criteria for the existence of a practical reason, suggesting it is in fact a counterfactual version of internalism. After all, instrumentalism holds that an agent has a reason to take whatever action she justifiably believes to be an optimal means to satisfying her most strongly desired ends. But instrumentalists allow for the possibility that the agent may fail to meet the criteria for having justified beliefs about her means. In such a case, the agent may not be motivated to take the optimal means to her ends. Even so, the instrumentalist will

normative judgments about practical reasons, even when these judgments are not treated as having truth-value.
insist that the agent *still* has a reason to follow the optimal means to her ends, and that she will not be fully rational if she fails to do so.

For example, suppose a woman smokes six packs a day because she wants the calming effect it has on her nerves. At the same time, she has an even stronger desire to maintain a healthy lifestyle. Unfortunately for the woman, evidence on the adverse health effects of smoking overwhelming suggests that these two desires are mutually exclusive—continued chain smoking likely leads to poorer health. If she knew the facts, she would see that she’d maximize the satisfaction of her desires if she stopped smoking, and she would quit. But let’s say the woman fails to inform herself, for either of two reasons. First, it might be that she unjustifiably believes a search for additional information would not serve her desires. Alternatively, she might be fully aware that additional information would improve her knowledge about how to satisfy her desires, but she can’t will herself to search for it. Either way, she persists in a state of ignorance, and she goes on puffing away. An instrumentalist analysis will suggest that the woman fails to act on justified beliefs about the optimal means to her ends. It supports the conclusion that the woman has a better instrumental reason to quit smoking than she does to continue. It does *not* suggest that the woman has no reason to quit smoking, or better reason to continue smoking, simply because she doesn’t give up her habit. Instead, the instrumentalist will maintain that the woman fails to act on the best of her reasons.

The challenge raised here asks why instrumentalism does not qualify as a counterfactual version of reasons internalism, since it attributes a reason to the smoking
woman based on what she would do, not what she actually does, if she were instrumentally rational. The answer is that this reasons attribution is still not contingent upon a desire that the woman would have if she were fully informed about the health effects of smoking. Even in her present state of ignorance, she by hypothesis does actually have the desire to maintain a healthy lifestyle. True, smoking is what the woman does, and this activity is caused by another desire to calm her nerves. But from this it does not follow that the woman lacks a desire to be healthy. Nor does it follow that her desire for health is weaker than her desire for the temporary stress relief afforded by smoking. Rather, by construction of the case, the woman does prefer health to temporarily calmed nerves, but due to her lack of relevant information, she fails to act on this preference. What this case suggests is that the desire whose presence is a necessary condition for an agent S to have a best reason R to A must exist as an element in S’s actual suite of desires. This condition explains why instrumentalism counts as an actual version of reasons internalism. The same condition can be found in the instrumentalist theories of Richard Brandt and Peter Railton. Both authors define one’s best reasons to act in terms of what would satisfy an agent’s desires, given that certain idealized conditions are satisfied—e.g., beliefs purged of false empirical information, errors in logic, and so on (cf. Hooker & Streumer, 2004, p. 62; Brandt, 1979, pp. 111 – 13; Miller, 2003, p. 208; Railton, 1986). Nevertheless, it is still the agent’s actual desires which fix the agent’s reasons for action. Indeed, for both Brandt and Railton, having full
information and flawless reasoning contribute to practical rationality precisely because these ideal conditions promote the satisfaction of the agent’s actual desires.

1.4.3 Williams’s actual internalism

Another widely recognized, but poorly understood version of actual reasons internalism can be attributed to Bernard Williams (cf. Williams, 1979, 1995a, 1995b, 2001). The earliest piece where Williams defends his reasons internalism is in the classic 1979 paper, “Internal and External Reasons.” There, Williams gives the following analysis of third-person reasons-statements of the form “S has a reason to A” (cf. Williams, 1979, in SLC, p. 292):

\[(W_0) \text{ If agent } S \text{ has a reason } R \text{ to do action } A, \text{ then if } S \text{ is practically rational, } S \text{ has some motivation } M \text{ which is served or furthered by } A\text{-ing.}\]

Because this is an early formulation of Williams’s view, I shall call it \((W_0)\). Williams says \((W_0)\) is the “internal” interpretation of “S has a reason to A.” As a terminological convenience, he uses the term “internal reason” to designate whatever falls into the extension of \(R\) in \((W_0)\). \((W_0)\) is clearly a variant of reasons internalism, because it takes the truth of a reasons-statement to be contingent upon the agent’s actual motivations, \textit{provided} that she is practically rational. According to Williams, an internal reasons-statement is falsified by the absence of some appropriate element from the agent’s \textit{subjective motivational set} (Williams, 1979, in SLC, p. 293). This is the set of
motivations that an agent actually has. Williams represents the subjective motivational set with the letter “S,” but to keep the notation employed in this chapter consistent, I shall use “\{M\}” to refer to the motivational set, and “M” to refer to some element in the set.

Some philosophers interpret Williams’s view as a competitor to instrumentalism (see especially Korsgaard, 1986; McDowell, 1998; Finlay, 2009). Williams takes pains to distance himself from a simple version of instrumentalism which he pejoratively calls the “sub-Humean model.” On this model, S has a reason to A if and only if S has some desire, the satisfaction of which S believes will be served by A-ing (Williams, 1979, in SLC, p. 292). Williams thinks this model is too crude, and he attempts to revise it in several ways. First, the sub-Humean model implies that S would have a reason to A even if she falsely believed that some desire of hers would be served by A-ing. Williams offers a well-known thought experiment to bring out the implausibility of this. We are to imagine a man who wants a gin and tonic. He believes that the gin bottle in his liquor cabinet contains gin, but unbeknownst to this man, his gin bottle is in fact filled with petrol. According to the sub-Humean model, this man would have reason to mix the contents of his gin bottle with the tonic and drink the concoction. Of course, this conclusion seems intuitively mistaken: it seems obvious that the man has no reason to drink petrol, particularly if he has a strong desire not to get sick!

14 Williams never discusses the possibility that motivations can change over time, and so it’s ambiguous whether the motivational set should be identified as the motivations an agent actually has at a particular moment, or throughout his or her lifetime. As a simplifying assumption, we can take the motivational set to be the set of motivations an agent has at a particular time t, where presumably this set includes motivations to act in the future.
In light of this case, Williams conjectures that a motivation $M$ will not give an agent $S$ a reason to $A$ if either $S$ has $M$ because of a false belief, or $S$ falsely believes that $M$ is served by $A$-ing (cf. Williams, 1979, in SLC, p. 293). In other words, in order for an agent to have practical reasons, his or her practical deliberations must be informed by true beliefs. This condition distinguishes Williams’s reasons internalism from the sub-Humean model, as well as the somewhat more sophisticated characterization of instrumentalism encapsulated in (I-1) through (I-5) in section 1.1. As we noted earlier with the case of the high-salt diet, (I-1) through (I-4) permit the possibility that an agent could falsely believe that $A$-ing is an optimal means to achieving one of her supremely desired ends, but still have a reason to $A$ provided that her means-ends belief is justified.

Unlike the sub-Humean model, though, the version of instrumentalism this chapter opened with does not necessarily support the claim that the man who wants a gin and tonic should have a reason to drink the petrol cocktail. In fact, the instrumentalist can deny that the man has any such reason by citing considerations similar to those adduced in the discussion of the chain-smoking woman. For instance, it could be argued that the man fails to have a justified belief about the petrol cocktail’s status as a means to his ends. He may be ignoring the marginal benefit of acquiring further information on which to base his means-ends beliefs. For the man would presumably maximize the satisfaction of his desires by avoiding the petrol cocktail, and the information required for him to see this is readily available. After all, if the man wanted to avoid sickness, he would just have
to take a whiff of the gas fumes curiously emanating from that gin bottle! In fact, Williams’s basis for “correcting” false beliefs is the same as the instrumentalist’s basis for ensuring justified beliefs. He assumes that “any rational deliberative agent has in his [motivational set] a general interest in being factually and rationally correctly informed” (Williams, 1995a, p. 37). This is secured by ensuring that the rational agent has true beliefs about the relation between an action he is considering and his motivational set (Williams, 2001, pp. 91 – 92). But whether or not agents can have practical reasons based on justified but false beliefs is an important point of disagreement between Williams’s internalism and instrumentalism.

Another way Williams’s internalism differs from the sub-Humean model, and supposedly more sophisticated instrumentalisms as well, concerns the nature of motivation. For instrumentalists, the type of motivational state which gives rise to practical reasons is desire. But for Williams, the motivational set \( \{M\} \) can include motivational states other than desires, such as “dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects, as they may be abstractly called, embodying commitments of the agent” (Williams, 1979, in SLC, p. 294).

---

15 An interesting variation on this thought experiment would assume that the man’s gin bottle was spiked with methyl alcohol. Methyl alcohol is lethal in small amounts. But it’s very difficult to differentiate from ethyl alcohol—the type found in alcoholic beverages. If the man’s cocktail had methanol in it, would he have a reason to drink it? Intuitively, we would think he does not (assuming he wants to live). The instrumentalist view can support this intuition just as well. Although it would be more difficult (and hence costly) to ascertain the true contents of his beverage, the man would presumably have much to gain from investing in additional information. Thus, if the man believed that there was nothing dangerous in his glass, and this belief was justified merely by the testimony of his four senses, the belief would be unjustified because it was not based on an optimal search of information. Thanks to Kevin Hoover for suggesting this adjustment on Williams’s thought experiment.
Williams’s interpreters—particularly those with anti-instrumentalist leanings—tend to see these remarks as a decisive indication that Williams’s view goes beyond any version of instrumentalism (see, for instance, Wong 2006, ch. 3; McDowell, 1998; Korsgaard, 1986; Finlay, 2009).

A third alleged difference between Williams’s internalism and instrumentalism concerns how the principles of practical reasoning are to be articulated. Looking back to the (RI) schema, the question is what principles, if any, should determine the practical reasons of an agent? What should count as \( P \)? We saw that instrumentalism cashes out \( P \) as the Instrumentalist Dictum. Judging from the clause in \( (W_0) \) which says that an agent’s having a practical reason to \( A \) depends on her having some motivation \( M \) which is “served or furthered” by doing \( A \), it’s tempting to think that Williams also endorses a means-ends conception of rational practical deliberation. But in fact Williams’s internalism does not limit practical rationality to means-ends reasoning. He says

\[
\ldots \text{there are much wider possibilities for deliberation, such as: thinking how the satisfaction of elements in [the agent’s motivational set] can be combined, e.g. by time-ordering; where there is some irresoluble conflict among the elements of [the motivational set], considering which one attaches most weight to (which, importantly, does not imply that there is some one commodity of which they provide varying amounts); or again, finding constitutive solutions, such as deciding what would make for an entertaining evening, granted that one wants entertainment (Williams, 1979, in SLC, p. 294).}
\]

Williams appears to be pointing to a capacity in rational agents to revise the contents of their motivational sets. It does sound like these operations belong to the domain of practical reasoning. However, it remains to be seen whether or not the sophisticated instrumentalist can incorporate them into the framework of means-ends reasoning. Some
instrumentalists have indeed argued that people are endowed with “second-order” or “maieutic” desires to choose or pursue some desires rather than others (cf. Frankfurt, 1971; Schmidtz, 2001). Whereas Williams sees the deliberative possibilities described above as foreign to means-ends reasoning, the instrumentalist could explain them in terms of the satisfaction of these higher-order desires. For instance, the choice of foregoing an opportunity to serve an immediate desire in order to fulfill another desire at a later time could itself be seen as serving a higher-order desire to ensure long-term satisfaction over the course of one’s whole life. In chapters 4 and 5, I defend a form of instrumentalism which holds that agents have higher-order desires which enable them to critique and revise lower-order desires.

Beyond what’s stated in the quoted passage, Williams is not much more specific about what he takes to be the principles which describe practical deliberation. Indeed, he seems to think that further specificity cannot be had:

…it is unclear what the limits are to what an agent might arrive at by rational deliberation from his existing [motivational set]…I regard it as a basically desirable feature of a theory of practical reasoning that it should preserve and account for that unclarity. There is an essential indeterminacy in what can be counted a rational deliberative process. Practical reasoning is a heuristic process, and an imaginative one, and there are no fixed boundaries on the continuum from rational thought to inspiration and conversion (Williams, 1979, in SLC, p. 297).

For Williams, an adequate account of the principles of practical rationality is inevitably open-ended, since there are many potential ways in which an agent’s activity may “served or furthered” by his or her \( M \). Indeed, Williams admits to regarding this vagueness as an advantage his theory holds over both instrumentalism and anti-internalist
views about practical rationality (Williams, Ibid.). In his subsequent writings on internalism, Williams pointedly refuses to add detail to the account of $P$. Instead, Williams deletes the locution “served or furthered” from later formulations of his view, which has the effect of removing any suggestion that he holds an instrumentalist picture of practical rationality. In “Further Notes on Internal and External Reasons” (2001), Williams gives his preferred way of expressing his position, which is the primary basis for the following formulation (cf. Williams, 2001, p. 91; see also Williams, 1995a, p. 35; McNaughton & Rawling, 2004, p. 117):

**Definition:** A sound deliberative route is an argument, or chain of reasoning, which is justified by some rational principle $P$ and contains only true premises.

(W₁) If agent $S$ has a reason $R$ to do action $A$, then if $S$ is practically rational,

(i) there is a sound deliberative route from $S$’s actual subjective motivational set, $\{M\}$, to $S$’s $A$-ing and

(ii) $S$ has $R$ because of (i).

Hereafter I shall treat (W₁) as Williams’s considered position. The “sound deliberative route” clause serves the same function as $P$ in the (RI) schema: both act as placeholders for some principle that justifies chains of reasoning that conclude with statements about an agent’s reasons for action. However, the sound deliberative route is
sound not only in the sense that it comports to principles of rationality, whatever those may be, but also in the sense that it proceeds from true premises. This added sense takes into account Williams’s restriction, adopted in light of the petroleum cocktail example, that reasons for action cannot be constituted by false beliefs.

Apart from the differences that set Williams’s (W₁) off from instrumentalism, there are important similarities as well. In particular, (W₁) is a version of actual reasons internalism. Williams’s writings leave little doubt that he would avow this classification. In “Internal Reasons and the Obscurity of Blame” (1995a), Williams tells us that an agent “has a reason to Φ only if he could reach the conclusion to Φ by a sound deliberative route from the motivations he already has” (Williams, 1995a, p. 35; emphasis added). Thus the “sound deliberative route” Williams speaks of seems to be justified by some procedural principle of practical rationality, which takes as basic inputs the agent’s actual motivations.

Another of Williams’s celebrated thought experiments provides a strong hint that he subscribes to actual reasons internalism.¹⁶ He asks us to consider the case of Owen Wingrave, a character in a short story by Henry James. Owen’s father urges him to join the army, since Owen’s filial duty is to carry on the Wingraves’ proud tradition of military service. Owen, however, is left completely cold by his father’s entreaties: in fact, “he hates everything about military life and what it means” (Williams, 1979, in SLC, p. 295). Williams notes that the elder Wingrave could have appealed to his son by saying he

¹⁶ Contrary to the way some have interpreted him (see Finlay & Schroeder 2008).
had a *reason* to join the army. And in saying this, Wingrave the elder may not intend to be making an internal reasons claim along the lines of \((W_1)\). Rather, he may insist that Owen has a reason to join the army, knowing full well that there was nothing in Owen’s motivational set that would lead, through practical deliberation, to his joining the army. In that case, Owen’s father would be making a statement about his son’s reasons which should be interpreted in what Williams calls “the external sense” (Williams, *Ibid.*). According to Williams, statements about what he calls “external reasons” do *not* treat the availability of a “sound deliberative route” from \(S\)’s motivational set to her \(A\)-ing as a necessary condition for \(S\) to have a reason to \(A\) (Williams, 1979, in SLC, p. 292; Williams, 1995a, p. 35). Understood this way, the elder Wingrave’s claim that Owen has a reason to join the army is not falsified by the absence of any actual motivation in Owen’s \(\{M\}\) to pursue a military career. Instead, Williams surmises, the elder Wingrave’s exhortations could be intended to make Owen aware of some consideration whose normative force he would appreciate, if only he were “considering the matter aright” (Williams, 1979, in SLC, p. 296). Thus Owen may neither recognize nor be motivated by a filial duty to join the army, and his father may be fully aware of all this. But the elder Wingrave’s point is that Owen has a duty to join the army, and by virtue of that duty alone he has a reason to join the army.

However, Williams argues that statements about external reasons fail to meet at least one truth condition that applies to *any* third-person reasons statement. I shall call this condition the *explanatory constraint on reasons*:
(E) If any agent $S$ has a reason $R$ to $A$, then it is possible that $S$ does $A$ because of $R$.

This proposition draws primarily on Williams (1995a, pp. 38-39). But (E) looms large in Williams (1979) as well. There, Williams draws attention to the fact that in ordinary discourse, actions are often explained by citing the reasons for which they are done. This is widely thought true of instrumental reasons, which figure in *reasons-giving* explanations such as the following (cf. Rosenberg, 1995, p. 37):

1. Given that it’s raining at time $t$, Smith’s strongest desire at $t$ is to stay dry.
2. Smith (justifiably) believes that carrying an umbrella is the optimal means for him to stay dry.
3. Because of (1) and (2), Smith would maximize the satisfaction of his desires by carrying an umbrella.

The Instrumentalist Dictum: For any action $A$, $S$’s best practical reason is to $A$ if and only if $S$ justifiably believes that $A$ maximizes the satisfaction of his or her desires.


5. Therefore, Smith carries an umbrella *because* his best instrumental reason is to carry an umbrella.
Smith’s desire to stay dry, together with his belief that carrying an umbrella is the best way to stay dry, constitutes an instrumental reason for Smith to carry an umbrella. And as suggested by the third premise of this explanatory argument, instrumental reasons can cause actions. Explanations of this kind are familiar in everyday life and in the social sciences. Max Weber believed that the explanatory principle stated in the fourth premise above has the status of a law, and therefore plays the same role in explanations of action as laws of physics or chemistry do in explanations of phenomena investigated in the natural sciences (Rosenberg, 1995, p. 33).

Of course, Williams is not committed to the idea that there all normative practical reasons are instrumental reasons. Nor is he committed to thinking that only instrumental reasons can figure in reasons-giving explanations. Williams’s point, as he expresses it, is rather that “the normative statement ‘He has a reason to Φ’ introduces the possibility of that reason being an explanation…if the agent accepts that claim (more precisely, if he accepts that he has more reason to Φ than to do anything else)” (Williams, 1995a, p. 39, my emphasis). Moreover, Williams insists that all normative practical reasons must be capable of explaining action, if there are any such reasons at all: “If there are reasons for action, it must be that people sometimes act for those reasons, and if they do, their reasons must figure in some correct explanation of their action” (Williams, 1979, in SLC, p. 293, emphasis added). These points are, to my eye, faithfully paraphrased by (E) above: if S has any (normative) reason R to do action A, then it must be possible that S
does *A for that reason*—in other words, *because* of *R*. On the other hand, if it’s not possible for any normative reason to explain action, it would follow from (E) that there are no normative reasons for action.

Williams goes on to voice skepticism that *external* reasons ever could fulfill the explanatory constraint. His argument furnishes insight into which type of modality he intends by the clause in (E) stating that it must be “possible” for someone to act *because* of her practical reasons. At various points, Williams suggest that the locution “possible” in (E) refers to what might be called *psychological possibility*—i.e., a possibility which is *consistent with the laws of psychology*. Ideally, the laws of psychology would be those articulated in a true and complete account of human psychology. Failing that, we will have to settle for the laws articulated by the *best current* theory of psychology. This interpretation of the possibility of a practical reason’s explaining an action is supported by Williams’s remarks about how reasons explain action. He says, in particular, that “nothing can explain an agent’s (intentional) actions except something that motivates him so to act” (Williams, 1979, in SLC, p. 295). But if a reason explains an action only by motivating it, then it’s *possible* for a reason to explain an action only if it’s *possible* for the reason to motivate the action. Motivation is a psychological concept, and whether or not some element of an agent’s psychology can provide a motivation is determined by the relevant psychological laws.  

So it is not possible—*psychologically possible*, we can

---

17 I shall use the term “motivation” to refer to something “internal” to an agent’s psychology which causes the agent to act. For instance, Smith’s belief that it’s going to rain and his desire not to get wet are both
say—for any agent $S$ to do $A$ because of practical reason $R$, unless it is consistent with the laws of psychology for $R$ to motivate $S$’s $A$-ing. Thus, Williams’s own comments, together with the logic of his argument for $(W_1)$, make it evident that he is committed to the following refinement on $(E)$:

$$(E-W) \text{ If any agent } S \text{ has a practical reason } R \text{ to } A, \text{ then (as a necessary condition) it is possible that } S \text{ does } A \text{ because of } R, \text{ where it is possible that } S \text{ does } A \text{ because of } R \text{ just in case it is psychologically possible for } R \text{ to motivate } S \text{'s } A\text{-ing.}$$

According to $(E-W)$, reasons for action must explain action in a manner which is permitted by the laws of psychology. Crucially, Williams has a very specific understanding of the psychological laws governing motivation. Recall that external reasons for action are not derived from the existing motivations of the agent. So if external reasons are to meet the explanatory constraint, the motivation to act on them cannot be found in an agent’s actual motivational set. Instead, Williams considers the hypothesis that a new motivation to act on the external reasons claim “$S$ has a reason to

items “internal” to Smith’s psychology. The combination of Smith’s belief (that it will rain) and desire (not to get wet) clearly plays a role in causing Smith to pick up his umbrella. Therefore, Smith’s belief-plus-desire is a motivation for Smith to pick up his umbrella. The rain clouds above Smith’s head, by contrast, are not “internal” to Smith’s psychology. So although the clouds may generate motivations for Smith, for example if he spots them when looking out his window, they are not themselves motivations for Smith. 
“may be generated in S the moment she comes to believe that claim (Williams, 1979, in SLC, p. 296). This hypothesis makes the truth of external reasons statements depend on what it means for one to come to believe that one has an external reason. As Williams understands them, “external reasons theorists” imply that one would come to believe in a true external reasons statement by “considering the matter aright”—that is, by engaging in successful practical deliberation. Thus, Williams reads external reasons statements “as roughly equivalent to, or at least as entailing, the claim that if the agent rationally deliberated, then, whatever motivations he originally had, he would come to be motivated to Φ” (Williams, 1979, in SLC, p. 296). In this quote, Williams suggests that external reasons theorists need to give a picture of practical reasoning according to which practical deliberation has the power to (1) yield conclusions about external reasons for action that do not derive from the agent’s actual motivations, and (2) to generate new motivations in the agent to act in accordance with the conclusions she draws about her external reasons. Part (1) of this picture represents the possibility of substantive practical reasoning, and part (2) represents the possibility of substantively rational agents capable of being motivated by the deliverances of substantive reasoning. The deliverances of substantive reasoning are, in Williams’s understanding, statements about external reasons. Thus, as Williams understands them, external reasons theorists can meet the challenge of explaining how external reasons satisfy (E-W) by showing that it’s compatible with the laws of psychology for people to be substantively rational agents.
As an illustration of the process of substantive practical reasoning that Williams has in mind, let us revisit the case of Owen Wingrave. Owen’s father might have supported the claim that Owen has reason to join the army with the following practical syllogism:

(1) One ought to do his filial duty.
(2) By joining the army, Owen Wingrave will do his filial duty.

*Substantive principle of practical rationality:* If one has a filial duty to do $A$, then one has a reason to $A$.

(3) Therefore, Owen has a reason to join the army.

The reason cited in (3) is, by Williams’s account, an “external” reason. Wingrave the elder concludes that Owen has this reason based on a substantive principle of practical inference which does not refer to any actual element of Owen’s motivation set. In particular, no mention is made of the fact that Owen has no desire to join the army. From Owen’s father’s perspective, what Owen wants to do is irrelevant. Instead, premise (1) of the argument posits an objective value—namely, that of doing one’s filial duty. The substantive principle licenses an inference from the fact that Owen ought to do his filial duty by joining the army to the conclusion that Owen has a *reason* to join the army. This is what it means for a substantive principle of reasoning to take an objective value as a “basic input.”
According to Williams, if the external reasons statement in the conclusion of this practical syllogism is to fulfill (E-W), then it should be psychologically possible that Owen comes to be motivated to join the army solely in virtue of being persuaded by his father’s argument. Furthermore, this motivation must have been generated after Owen comes to believe the conclusion through a process of substantive practical reasoning, and it could not have been present among Owen’s existing motivations prior to his hearing the argument. Yet Williams doubts that new motivations can be generated in this way—that is, via rational deliberation alone. To appreciate Williams’s skepticism about external reasons, it helps to consider Hume’s theory of motivation, or HTM. My summary of this theory follows the one set down by Christine Korsgaard in “Skepticism about Practical Reason” (1986) (cf. Korsgaard, 1986, in SLC, p. 300, 301):

*Hume’s theory of motivation (HTM)*

(H1) No operation of reason is a passion (desire).

(H2) Only the passions (desires) are sufficient to produce motivations.

(H3) The only operation of reasoning that transmits motivational force is reasoning about the optimal means to satisfying a passion (desire).

Williams does not endorse every claim of HTM. However, in “Internal and External Reasons,” he does express sympathy with another claim which HTM arguably

---

18 I also follow Robert Audi (2006, ch.2) in glossing what Hume calls “passions” as desires.
implies: namely, that reasoning alone cannot “give rise” to a motivation (Williams, 1979, in SLC, p. 296). This thesis is entailed by the propositions (H1) and (H2) of HTM.

According to (H2), only the passions (or desires) are sufficient to generate motivation. But if, as (H1) indicates, the passions (desires) are distinct from operations of reason, then no operation of reason is a sufficient source of motivation. (On the other hand, (H1) does not deny that operations of reason can be necessary sources of motivation.) Therefore, we can validly infer the following conclusion:

(H) Reasoning alone is not sufficient to produce a motivation.

Williams is plainly convinced that (H) is true. And it’s on the basis of this thesis that he suspects external reasons statements must be false. If external reasons statements are to satisfy Williams’s explanatory constraint, then they must motivate purely by virtue of sound practical reasoning, and independently of one’s existing motivations. When an agent is motivated by an external reason she believes she has, then “ex hypothesi, there is no motivation for the agent to deliberate from, to reach this new motivation” (Williams, 1979, in SLC, p. 296). Williams declares that he sees “no reason to suppose that these conditions could possibly be met” (Ibid.). In effect, he suggests that it is not psychologically possible for an agent to A because he or she has an external reason to A. According to (H), it is not permitted by the laws of psychology for an agent to arrive at a
motivation to $A$ from a sound deliberative route that does not “start from” the agent’s existing motivational set.

Williams’s skepticism that external reasons statements could meet the explanatory constraint follows from his seeing “great force in Hume’s basic point,” namely (H) (Ibid.). If external reasons statements meet (E-W), then it must be possible for them to motivate solely in virtue of the agent’s coming to believe them through sound practical reasoning. But given that (H) is true, external reasons statements cannot so motivate, and therefore they fail to meet (E-W). On the strength of these considerations, Williams concludes that “it is very plausible to suppose that all external reason statements are false” (Ibid.).

Although it’s evident that Williams accepts (H), he is by no means committed to any particular claim in HTM. We’ve already seen that Williams does not think practical reasoning is limited to instrumental, or means-ends reasoning. He would thus reject (H3). And because he thinks a greater diversity of motivations can produce action than merely desires alone, he would have to deny (H2). Nevertheless, Williams can retain his commitment to (H). There are other possible claims that Williams could hold which imply (H), but which are still consistent with the comments he makes to distance his internalist view from the sub-Humean model. First, Williams could hold the thesis that no operation of reasoning is an ultimate source of motivation. An “ultimate” source of motivation is one that has motivational force intrinsically, and not in virtue of any other source of motivation. This thesis, which we can call (W1), allows for the existence of
ultimate sources of motivation other than desires. Meanwhile, (W1) excludes practical reasoning itself from the ultimate sources of motivation.

(H) says that reasoning alone is not sufficient to motivate, but it doesn’t follow from this that operations of reason have no motivational force at all. Rather, Williams can adopt a thesis, similar to (H3), which holds that some operations of reasoning can have motivational force by “transmitting” it from some ultimate source of motivation. These motivationally efficacious operations of reasoning need not be limited to means-ends reasoning. Williams suggests, both in his formulations of his reasons internalism and in his arguments for it, that the ultimate sources of motivation from which reasoning can transmit motivational force are to be found within the agent’s motivational set. As I’ve shown, Williams’s classic argument for (W1) concludes with the assertion that there are no external reasons for action.

On the other hand, Williams obviously holds that internal reasons for action do exist. These contrasting verdicts about reasons hinge entirely on Williams’s explanatory constraint (or E-W): external reasons do not exist because they fail (E-W), whereas internal reasons exist because they satisfy (E-W). Williams denies that external reasons exist because, in light of (H), it is not psychologically possible for them to motivate action. By contrast, Williams’s conclusion implies that internal reasons meet (E-W) because it is psychologically possible for them to motivate. By definition, internal reasons are derived through a “sound deliberative route” from the actual motivations of the agent. This “sound deliberative route” is a process of procedural practical deliberation
corrected for any false premises. The motivational efficacy of internal reasons seems to be founded in the fact that procedural operations of practical reasoning “start from” the agent’s actual motivations. It thus appears that Williams inherits from Hume a “hydraulic” conception of the relationship between motivation and practical reason. On this conception, operations of reason are insufficient to produce a motivation, but they can channel motivational force from its ultimate sources in an agent’s motivational set, thereby making action conform to the dictates of practical rationality.

Thus, the following analogue to Hume’s (H3) is implicit within Williams’s argument: procedural practical reasoning is the only operation of reasoning that transmits motivational force. This thesis we can call (W2), which in addition to (W1) comprises Williams’s theory of motivation (or WTM):

\[ \text{Williams’s theory of motivation (WTM)} \]

(W1) No operation of reason is an ultimate source of motivation.

(W2) The only operation of reasoning that transmits motivational force is procedural practical reasoning.

(W1) and (W2) imply (H), the thesis that reasoning alone is not sufficient to produce a motivation. (W1) rules out the notion that practical reasoning can be an ultimate source of motivation. But insofar as practical reasoning can motivate, its motivational force has to come from somewhere. (W2) tells us that practical reasoning
motivates only by channeling or “transmitting” motivational force from the agent’s existing set of ultimate motivations. These existing motivations will be elements of the agent’s motivational set, or \( \{M\} \). Of course, both HTM and WTM are sets of empirical claims about the nature of human psychology. Perhaps further empirical investigation will reveal one or both to be false, but the Williams-style reasons internalist will go to the mat to defend WTM. In chapter 3 (section 3.3 – 3.5), I shall discuss several theoretical considerations that support a presumption in favor of WTM.

One more facet of Williams’s internalism has to be explained. Clause (ii) in \((W_1)\) says that \(S\) has a practical reason to \(A\) because there is a sound deliberative route from \(S\)'s motivational set to \(S\)'s \(A\)-ing. For Williams, an agent’s practical reasons exist in virtue of the presence of a sound deliberative route from the agent’s motivations. This much can be discerned from (E-W) and WTM. Practical reasons, if they exist, must be capable of motivating agents in order to explain action. And since practical reasoning alone is insufficient to generate motivation, their motivational efficacy must derive from the agent’s existing motivations. Thus, practical reasons are explained by an agent’s motivational set combined with sound practical reasoning. We have reasons for action because we have the motivations we do, and because we can deliberate soundly about them.

To conclude my discussion of Williams’s brand of internalism, I will re-formulate his view to fit the standard way in which reasons internalism has been expressed elsewhere in this chapter. In his 2001 postscript, Williams states his position this way: for
any agent, $S$, “[$S$] has a reason to [$A$] only if there is a sound deliberative route from [$S$’s] subjective motivational set...to [$S$’s $A$-ing]” (Williams, 2001, p. 91). It’s been established that by a “sound deliberative route” from the agent’s $\{M\}$, Williams is likely referring to an operation of procedural practical reasoning which is free of false premises.

Furthermore, given that this sound deliberative route ends with the agent doing what he or she has reason to do, we can infer that the agent in question is procedurally rational. The assumption supporting this interpretation is simply that an agent’s $A$-ing is caused by an effective motivation to $A$, and if an agent has an effective motivation to do as a sound process of procedural reasoning concludes he has (best) reason to do, then he is procedurally rational. Therefore, we have the following reconstruction of Williams’s actual reasons internalism:

(W) If agent $S$ has a reason $R$ to do action $A$ in circumstances $C$, and $S$’s beliefs about $C$ are true, then if $S$ is procedurally rational,

(i) $S$ has some motivation $M$ to $A$ and

(ii) $S$ has $R$ because of (i).

We’re now in a position to summarize Williams’s argument for (W):

*Williams’s Classic Argument for Reasons Internalism (from Williams 1979)*

(1) Either there are internal reasons for action or there are external reasons (or both).
(2) By (E-W), if there is an external reason $R^e$ to $A$, then it is psychologically possible for $R^e$ to motivate $S$’s $A$-ing.

(3) But if it is psychologically possible for $R^e$ to motivate $S$’s $A$-ing, then substantive practical reasoning alone is sufficient to produce a motivation in $S$.

(4) But by (H), practical reasoning alone is not sufficient to produce a motivation.

(5) Hence, there are no external reasons for action.

(6) Therefore, there are only internal reasons, i.e., practical reasons of which (W) is true.

This depiction locates the master strokes of Williams’s argument at premises (2) and (4). In line(2), Williams applies (E-W) as an ontological criterion which external reasons must satisfy. What is required is the psychological possibility that external reasons are sufficient to motivate agents. Williams entertains the hypothesis, expressed in line (3), that external reasons can motivate solely in virtue of a process of substantive practical reasoning. But in line (4), we see that substantive practical reasoning alone cannot motivate if the *Hume-Williams thesis*, (H), is true. Thus, external reasons fail to meet (E-W), and so there can be no external reasons. Internal reasons, on the other hand, do meet (E-W) because they issue from *procedural* practical reasoning. Procedural practical reasoning establishes the normative reasons of an agent through a series of inferences from the agent’s actual motivations. Procedural reasoning can motivate agents to act on their reasons by transmitting motivational force from the agent’s existing
motivational set. Thus, according to the above picture, Williams’s entire edifice rests on an ontological claim—namely, (E-W), which draws a necessary connection between the existence of practical reasons and the psychological capacities of agents, and a psychological claim—captured in (H)—which implies that only internal reasons derived from procedural practical reasoning are compatible with the ontological claim. In the following sections, I will survey rival theories of practical rationality which object to Williams’s view either by rejecting (H) (section 1.5) or (E-W) (section 1.6).

1.5 Counterfactual reasons internalism: Practical reasoning beyond the motivational set?

Some philosophers are willing to throw Williams’s psychology into question while remaining faithful to his explanatory constraint on reasons, (E-W). Williams applies the adjective “external” to any reasons claim which doesn’t derive from some actual motivation in the agent. But this is a bit of a misnomer, since it is possible for a reasons statement to meet all the conditions laid out in (RI) and (E-W) without involving the agent’s actual motivational set. Christine Korsgaard shows how this is possible in “Skepticism about Practical Reason” (1986). Korsgaard’s position is indeed another variant of reasons internalism, but it is an internalism of the counterfactual type.

Korsgaard is careful to point out that she has no quarrel with Williams’s explanatory constraint. She states, first of all, that “[p]ractical-reason claims, if they are really to present us with reasons for action, must be capable of motivating rational
persons” (Korsgaard, 1986, in SLC, p. 303). Korsgaard’s name for this truth condition on reasons claims is “the internalism requirement.”

It’s evident from this passage that Korsgaard’s internalism requirement is not very different from (E-W). A capacity to be motivated by a practical reason would certainly make it possible for a rational agent to carry out the act favored by the reason. As it’s expressed, Korsgaard’s requirement refers to rational agents, whereas (E-W) refers broadly to any agent. Importantly, however, Korsgaard incorporates within her definition of a rational agent a capacity to be motivated by the operations of successful practical reasoning. She characterizes a practically rational agent as one who is “not merely capable of performing certain rational mental operations, but capable also of transmitting motive force, so to speak, along the paths laid out by those operations” (Korsgaard, 1986, in SLC, p. 304). Thus, like (E-W), Korsgaard’s requirement restricts the principles of practical reasoning that there are—\( P \) in (RI)—to the space of psychological possibility. For Korsgaard, an agent \( S \) is practically rational if and only if \( S \) can deliberate in accordance with \( P \), and it’s psychologically possible for \( S \) to act in accordance with \( P \).

The focal point of the disagreement between Korsgaard and Williams is over the motivational psychology of practical reasoning. Recall from section 1.4.3 that Williams’s theory of motivation, or WTM, implies the Hume-Williams thesis, (H). (H), in turn, implies that it is not psychologically possible for substantive practical reasoning to motivate agents. If substantive reasoning could motivate, it would have to be an ultimate source of motivation sufficient to move agents by itself. But (H) asserts that no operation
of reasoning is by itself sufficient to motivate. Korsgaard, however, challenges Williams’s account of the link between motivation and practical reasoning from a Kantian vantage point. She wants to argue that “there are operations of practical reason which yield conclusions about actions and which do not involve discerning relations between passions (or any pre-existing sources of motivation) and those actions” (Korsgaard, 1986, in SLC, p. 301). Here Korsgaard seems to be describing operations of substantive practical reasoning. But since she is committed to (E-W), the practical reasons delivered by substantive reasoning must be able to motivate rational agents. Thus Korsgaard’s main burden is to give an account of how substantive practical reasoning can motivate. This of course requires an argument against WTM, and another in favor of a psychology according to which substantive reasoning alone can have motivational effects. Developing such an alternative is not Korsgaard’s immediate concern in “Skepticism…,” but she does describe what its general features might be and what consequences it would have for (E-W).

Korsgaard suggests that if Williams’s explanatory constraint can be cut loose from WTM, it can be interpreted anew. She writes:

If one accepts the internalism requirement, it follows that pure practical reason will exist if and only if we are capable of being motivated by the conclusions of the operations of pure practical reason as such. Something in us must make us capable of being motivated by them, and this something will be part of the subjective motivational set (Korsgaard, 1986, in SLC, p. 308).

Korsgaard tells us that Kant himself took on this task in the Third Section of the *Foundations* (Korsgaard, 1986, in SLC, p. 310).
Korsgaard’s take on Williams’s explanatory constraint (or, the “internalism requirement”) comes to this: substantive practical reasoning could satisfy (E-W) if every *rational* being has a disposition to be moved by it as an element of their motivational sets. The possibility adverted to, here, is one where a disposition to be motivated by a substantive principle of practical reasoning itself belongs to the rational agent’s motivational set. If that is so, then any practically rational agent will deliberate in accordance with this principle *and* be motivated by the conclusions it yields about what there is reason to do. Korsgaard has thus envisioned a way in which an “external” reason for action can satisfy the conditions on practical reasons expressed in (RI), as well as Williams’s explanatory constraint (E-W). This places her view securely within the reasons internalist camp.

Another facet of the psychological dispute between Korsgaard and Williams concerns the explanatory direction of the relationship between practical reasons and motivation. For Williams, an agent has a practical reason to *A* *because* she would arrive at a new motivation to *A* via a sound deliberative route from her existing motivations. By contrast, Korsgaard’s internalism makes practical reasons explanatorily prior to motivation. For Korsgaard, it’s *because* we have certain practical reasons that, provided we are practically rational, we would be motivated to act in a manner favored by those reasons. Korsgaard suggests as much when she cites Nagel (1970) as arguing that “investigations into practical reason will yield discoveries about our motivational capacities” (Korsgaard, 1986, in SLC, p. 309). Both Korsgaard and Nagel agree with
Williams that (E-W) is a constraint on reasons: that is, if practical reasons exist, then they must be capable of motivating agents. However, Korsgaard and Nagel part with Williams in thinking that the motivational efficacy of practical reasons emanates from practical reasoning by itself. Where Williams and Korsgaard disagree is on the truth of (H), which is a consequence of WTM. Williams holds that the only principles of practical reasoning capable of motivating agents are procedural, because procedural principles derive their motivational force from the agent’s existing motivations. Korsgaard and Nagel, by contrast, hold that substantive principles of practical reasoning are capable of motivating, provided that the agent has a disposition to act in accordance with the conclusions of a process of substantive practical deliberation.

Korsgaard’s account of how substantive reasoning can satisfy (E-W) is what most clearly makes her a counterfactual reasons internalist. On her view, the actual motivations of agents don’t determine what practical reasons they have. Instead, one’s reasons for action are specified by substantive principles of practical rationality. In other work, Korsgaard adds some detail to what she thinks the principles of practical reason consist in (Korsgaard, 1996, ch. 3). Her account begins with an important phenomenological observation concerning the experience of choice and practical deliberation. Choice and deliberation, Korsgaard tell us, involve the capacity to reflect upon our own desires, impulses, and other motivations (Korsgaard, 1996, in SLC, p. 94, 96). Upon perceiving cloudy skies, one might experience an impulse to believe it will rain. Despite this impulse, one can still ask whether there is a reason to believe it will
rain, and if one concludes that there is not, one can resist the original impulse. Similarly, upon perceiving cloudy skies, one might experience an impulse to take out an umbrella, but again, one can still ask whether this impulse constitutes a reason to act in that way. And if not, one can act otherwise. Korsgaard refers to this capacity to distance oneself from one’s own motivations and subject them to scrutiny as a power of self-consciousness (Korsgaard, 1996, in SLC, p. 96).

According to Korsgaard, human beings are uniquely endowed with the capacity for self-consciousness (Korsgaard, 1996, in SLC, p. 94). And it’s through self-consciousness that we engage in practical deliberation. Our practical deliberations, in turn, are governed by a “principle or law” that we take to be expressive of our own practical identities. Korsgaard defines practical identity as “a description under which you find your life to be worth living and your actions to be worth undertaking” (Korsgaard, 1996, in SLC, p. 96). Our practical identities provide a source not only for our practical reasons, but also our moral obligations. To neglect either threatens the integrity of our practical identities, because it undermines our ability to identify ourselves with a description of our lives and actions that we value. We all have reason to act in a way that maintains the integrity of our own practical identities, because it is in virtue of our practical identities that we can discern our reasons for action through practical deliberation. In effect, every agent has a reason to maintain his own practical identity, because one’s practical identity is what makes one the sort of creature that can have any reasons and values at all.
But Korsgaard thinks there are constraints on the way in which any rational agent should conceive of his or her practical identity. Most importantly, all rational agents should recognize themselves as human beings. For the human being is “a reflective animal who needs reasons to act and to live” (Korsgaard, 1996, in SLC, p. 103). Unless one was human, one would not be capable of forming a practical identity. Since being human is “the source of all reasons and values,” Korsgaard argues, any agent who values his or her own practical identity rationally ought to value the condition of possibility for that identity—namely, one’s humanity. But then, in valuing themselves as human beings, rational agents ought also to value humanity “for its own sake” (Korsgaard, 1996, in SLC, p. 103, 104; p. 106; see also Korsgaard, 1996, ch. 4). Korsgaard concludes that any rational agent must have in her motivational set a disposition to abide by Kant’s “moral law,” also known as the Kingdom of Ends formula of his Categorical Imperative. This principle demands that we “act only on maxims that all rational beings could agree to act on together in a workable cooperative system” (Korsgaard, 1996, in SLC, p. 95). When all is said and done, Korsgaard contends that all rational agents have a reason, and therefore a motivation, to value the humanity in themselves and in others. She treats Kant’s moral law is a substantive principle of rationality, and she suggests that any rational agent must have a disposition to follow the moral law as an element in his or her motivational set. The moral law is a substantive principle, not a procedural principle, because the practical reasons it provides are not derived from the actual motivations of any particular agent. Rather, the “input” to the moral law is the objective value of
humanity, and its “outputs” are practical reasons to act on maxims that could be agreed upon by cooperating rational agents. Of course, (E-W) requires that any reason derived from the moral law must be capable of motivating the relevant agent. But Korsgaard argues that the moral law can satisfy (E-W) if a disposition to abide by its edicts is part of the motivational set of the agent. Even if a disposition to follow the moral law is included in the agent’s \(\{M\}\), that does not make it a procedural principle of rationality, because it remains the case that the reasons it provides are derived from the value of humanity, and not from any particular agent’s actual motivations.

Korsgaard never rules out the possibility that an agent may fail to be motivated by a concern for the humanity of others, as Kant’s moral law obliges us to do. But in that case, the agent would not be fully rational. For then she would fail to value a pre-condition of her own practical identity—her humanity—and so she would fail to value something which makes possible her capacity to value anything at all. On the other hand, a counterfactual internalist like Korsgaard holds the following conditional to be true: if an agent is practically rational, then she would be motivated by the conclusions of substantive practical reasoning because she would have a disposition to act on substantive reasoning as part of her motivational set. Like Kant himself, Korsgaard holds that a capacity to be moved by the moral law is a requirement of rational agency.

Ultimately, the difference between actual and counterfactual internalism amounts to this: for actual internalists, principles of practical reasoning do not place determinate normative constraints on what motivations must be contained within an agent’s
motivational set. At most, operations of practical reason rank or re-organize existing elements of \( \{M\} \) in the ways that Williams suggests (time-ordering, resolving motivational conflicts, prioritizing, etc.). For counterfactual internalists, by contrast, principles of practical of reason can place determinate normative constraints on the contents of an agent’s motivational set. Elements such as a concern for others would have to be present in the agent’s motivational set, if the agent were rational.

On the other hand, there is a sense in which all reasons internalisms are counterfactual internalisms. All forms of internalism require that the rational agent has a disposition to be motivated by the principles of practical rationality. Real-life agents may lack such a disposition, but insofar as they lack it, the reasons internalist would diagnose them as not fully rational. In this way, all internalists recognize the phenomenon of akrasia, or weakness of will. Korsgaard’s view suggests that an agent’s lacking a disposition to be motivated by principles of substantive practical reasoning, such as Kant’s moral law, implies that the agent is not fully rational. But it turns out that the actual internalist is committed to something similar: if one lacks a disposition to be motivated by principles of procedural reasoning, such as the Instrumentalist Dictum, then she is not fully rational. Korsgaard contends that flesh-and-blood agents are susceptible to what she calls true irrationality, which is an inconsistency among one’s own intentional attitudes. True irrationality involves a failure to be convinced or motivated by a consideration that you yourself believe you rationally ought to be convinced or motivated by. A host of factors can cause this: Korsgaard lists rage, passion, depression,
distraction, and mental or physical illness (Korsgaard, 1986, in SLC, p. 304). As an example of true irrationality, Korsgaard cites the possibility that someone could know that an action is an optimal means to a desired end, and yet be utterly unmoved to carry the action through (Korsgaard, 1986, in SLC, p. 303). This would constitute a failure to be moved by the Instrumentalist Dictum. In cases such as this, the instrumentalist would diagnose the agent as being irrational (compare I-3 above). Even for the instrumentalist, then, there is at least one motivation that any rational agent must have in his or her \(M\). This is the disposition to act as the Instrumentalist Dictum requires.

Similar comments apply more generally to any actual reasons internalist: an agent \(S\) is practically rational just in case \(S\) has within his or her motivational set the disposition to act according to procedural principles of practical rationality. The reasons of an agent are then assessed relative to the motivational set that she would have, if she were procedurally rational. However, since the actual internalist recognizes only procedural operations of practical rationality, this counterfactual condition on rationality is weak compared to the conditions imposed by counterfactual internalists. Procedural principles are functions from the agent’s actual motivations to her reasons. Consequently, procedural principles will generate different reasons for agents with different motivational sets. Under actual internalism, then, rational agents may have different

\[^{20}\text{Instrumentalists need not deny the possibility of instrumental irrationality, as Hume may have (cf. Korsgaard, 1986, in SLC, p. 303). Indeed, they would be wisest not to, given widespread empirical evidence that people do display, and even acknowledge, failures to take the best means to maximizing the satisfaction of their own desires.}\]
reasons if their motivational sets are different. Counterfactual internalists, by contrast, recognize substantive principles of rationality. Substantive principles guarantee that there are some reasons for action that all rational agents have—for example, reasons to be concerned with the humanity of others. This is because substantive principles of rationality do not make one’s practical reasons in any way dependent upon one’s actual motivations. Instead, substantive principles derive practical reasons from objective values.

1.6 Reasons Externalism: Practical reasoning beyond the explanatory constraint?

We’ve just seen that Williams’s argument for actual reasons internalism can be challenged at the fourth premise. A counterfactual reasons internalist could argue, as Korsgaard does, that agents can be motivated by the conclusions of substantive practical reasoning. If so, then “external reasons,” as Williams calls them, would satisfy Williams’s explanatory constraint, (E-W). Another tack in undermining William’s argument would be to reject the second premise. One could deny that normative reasons for action, if they exist, must satisfy either version of the explanatory constraint, be it (E) or (E-W). This position can most appropriately be called reasons externalism, since on this view practical reasons are conceived of as being “external” to any motivation a particular agent has or even would have in virtue of practical reasoning. I capture the general sense of reasons externalism in the following proposition:
(RE) Agent $S$ has a practical reason $R$ to do act $A$ in circumstances $C$ if and only if some ideally practically rational reference agent, $S'$, has $R$ in $C$.

Internalists and externalists differ on two dimensions. Unlike actual reasons internalists, reasons externalists take practical reasons to be based on objective values, such as the well-being of oneself or of others, and not on the actual motivations of any particular agent. Counterfactual reasons internalists agree with the externalists on this score. But in addition, reasons externalists differ from all reasons internalists in denying that the explanatory constraint, at least as it’s articulated in (E) or (E-W), is a condition on the existence of normative practical reasons. The more general formulation of the explanatory constraint, (E), says that if agent $S$ has a reason $R$ to do $A$, then it’s possible that $R$ explains $S$’s $A$-ing. However, some reasons externalists would argue that a reason for action can apply to an agent even if it’s not possible for any of that agent’s actions to be explained by that reason.

For instance, suppose Isabelle knows that taking the medicine her doctor prescribes would cure her of a serious illness. Isabelle, however, has no motivation whatsoever to get better. She prefers to stay sick. In fact, let’s up the ante by assuming that Isabelle suffers from some psychological defect which makes it impossible for her to acquire any motivation to take her medicine. Any internalist philosopher who accepts (E-W) will have to conclude that Isabelle has no reason to take her medicine. That said, actual internalists and counterfactual internalists may explain Isabelle’s lack of a reason
differently. An instrumentalist, for instance, would tell us that Isabelle lacks a reason because she has no desire to be healthy in the first place, and so she cannot have an instrumental reason to pursue the means to get healthy. From an instrumentalist point of view, Isabelle can lack a reason to take her medicine, but still be practically rational: she can act on all her best instrumental reasons, while not having an instrumental reason to take her medicine.

A counterfactual internalist, by contrast, might hold that one’s health and well-being is an objective value worth pursuing for its own sake. In that case, a substantive principle of practical reasoning will guarantee any rational agent a reason—indeed, probably a best reason—to take his or her medicine. But if Isabelle cannot possibly be motivated to take her own medicine, then she is not a rational agent. However, if (E-W) is true, Isabelle does not fail to be a rational agent in the sense that she doesn’t act on her best reasons. Instead, she fails to be rational in the sense of not having any reasons for action, at least with respect to her health. She is non-rational about matters relevant to her health, because she is unresponsive to health-related considerations that would motivate a rational agent. Thus, for the counterfactual internalist, Isabelle lacks a reason to take her medicine, but she cannot be practically rational. Isabelle lacks a reason to take her medicine in the same way a plant lacks practical reasons to do anything.

A reasons externalist, however, would read the case differently: Isabelle both has a reason to take her medicine, and she is potentially irrational in the sense that she fails to act on this reason. The externalist may agree with the counterfactual internalist that
one’s own health is intrinsically valuable, and that this guarantees for any rational agent a reason—again, perhaps even a best reason—to take one’s medicine. However, because the externalist rejects (E-W), the impossibility of motivating Isabelle to take her medicine would not imply that she doesn’t have a reason to do so. Instead, Isabelle has a reason she fails to act on. And if this reason is her best reason, then Isabelle is either genuinely irrational, or she is at least open to rational criticism.

As Isabelle’s case suggests, the externalist does not condition any agent’s reasons on the possibility that her actions can be explained by them. Instead, one’s reasons are determined with respect to a reference agent. The reference agent is represented in (RE) as $S'$. Different externalists specify the characteristics of the reference agent in different ways, but the essential feature of $S'$ is that he or she is ideally practically rational. According to reasons externalism, any actual agent $S$ has a reason to do whatever $S'$ would do. If (E) were true, there would have to be some way for facts about what $S'$ would do under a given set of circumstances to explain $S$’s actions under the same circumstances. And if (E-W) were true, there would have to be a psychological mechanism which motivates $S$ to act in ways that $S'$ would act under the same circumstances. Actual internalists characterize this mechanism as an operation of procedural practical reasoning, and counterfactual internalists characterize it as an operation of substantive practical reasoning (or as operations of substantive and procedural reasoning working in tandem). But since the reasons externalist rejects both (E) and (E-W), there need not be any possible way to explain $S$’s actions on the basis of
what $S'$ would do. Solely the fact that $S'$ would act in certain ways is sufficient to support real-life agents’ reasons to act in those ways. In this way, externalism allows for cases in which a reason cannot possibly explain the actions of the agent who bears that reason.

However, externalists can accommodate a more qualified explanatory constraint on practical reasons. Derek Parfit, for instance, suggests that even though it might be impossible for an agent like Isabelle to be motivated by a reason to take her medicine, this reason can still have explanatory potential in the uncontroversial sense that “[p]eople sometimes take the medicine that they know they need” (Parfit & Broome, 1997, p. 114, n. 28). With this remark, Parfit seems to be proposing an alternative explanatory constraint on reasons—one which is much weaker than (E-W) or even (E). (E) and (E-W) index the explanatory potential of a practical reason to particular agents said to have the reason. Thus, Isabelle has a reason to take her medicine only if it’s possible for her to take her medicine because of this reason. By contrast, Parfit’s qualified explanatory constraint suggests that Isabelle has a reason to take her medicine only if it’s possible for some reference agent to take her medicine. In Isabelle’s case, the reference agent is some other person who is rational enough to be motivated to take the medicine he or she needs.

Different externalists defend different characterizations of the relationship between the reference agent and actual agents. Parfit himself suggests the fairly weak restriction that $S'$ has to be sufficiently similar to $S$. For example, people who are motivated to take the medicine they need are like Isabelle at least insofar as they are persons. Abstractly, we can think of the similarities between $S$ and $S'$ as placing them in
the same *class* of agents, \{S\}. With these considerations in mind, we can formally state Parfit’s qualified explanatory constraint on reasons as follows:

\( (E-P) \) If \( S \) has a practical reason \( R \) to \( A \), and if there is some reference agent \( S' \) such that \( S \) and \( S' \) are both members of the same class of agents, \( \{S\} \), then it is possible that \( S' \) does \( A \) *because of* \( R \).

Parfit is aware that \( (E-P) \) requires something very different from both \( (E) \) and \( (E-W) \). He notes that philosophers committed to the latter claims would insist that “for it to be true that some particular person has some normative reason [for action], it must be possible that, on *this* occasion, and *without* any further change in this person’s motivational state, *this* person should act for this reason” (Ibid.). His reply, though, is that conceptualizing the explanatory constraint along these lines would only make Williams’s argument for internalism beg the question. For then the explanatory constraint would simply be a restatement of Williams’s internalism, \( (W) \), and therefore it couldn’t enter into an argument for \( (W) \). Russ Shafer-Landau lodges a similar charge.\(^{21}\) He formulates

\(^{21}\) In fact, the question-begging charge is widely voiced by Williams’s critics. Stephen Finlay deploys several variations of it in Finlay (2009). The variation I defend Williams against here is the most common one. The accusation is that Williams sets up his explanatory constraint on reasons, but then *defines* the “possibility” of being motivated by a reason as the possibility of being motivated by procedural practical reasoning. If that were the case, Williams’s argument would indeed beg the question, since reasons internalism asserts that the only practical reasons are ones which can motivate the agent via procedural reasoning. However, these critics fail to see that Williams’s explanatory constraint makes no particular claims about *how* practical reasons motivate. It claims only that practical reasons, if they exist, have to motivate in *some* way or another. The claim that only procedural reasoning is sufficient to motivate is
Williams’s reasons internalism as the following (Shafer-Landau, 2003, in SLC, pp. 312-313):

(A) Necessarily, if $S$ has a good reason to $\Phi$ at $t$, then $S$ can be motivated to $\Phi$ at $t$.

Shafer-Landau goes on to interpret the modal locution “can be motivated” as follows:

(B) One can be motivated to $\Phi$ (i.e., it is possible for one to be motivated) to $\Phi$ if and only if one is already motivated to $\Phi$, or one would be motivated to $\Phi$ after rationally deliberating from one’s existing motivations.

To say, as (B) does, that one would be motivated to $\Phi$ after rationally deliberating from one’s existing motivations is to say that one would be motivated to $\Phi$ after concluding that she should $\Phi$ by a procedural operation of practical reasoning. So, (A) and (B) together express Williams’s reasons internalism. Shafer-Landau then quotes Williams’s statement of his explanatory constraint in Williams 1995a (p. 39): “If it is true that $[S]$ has a reason to $\Phi$, then it must be possible that he should $\Phi$ for that reason; and if introduced separately as the Hume-Williams thesis, (H). (E-W) and (H) together imply Williams’s reasons internalism, or (W).
he does act for that reason, then that reason will be the explanation of his acting.” But the kind of possibility Shafer-Landau takes Williams to mean by “it must be possible…” is that expressed in (B) (Shafer-Landau, 2003, in SLC, p. 314). Interpreted this way, Williams’s explanatory constraint simply re-expresses reasons internalism, and so Williams’s argument for reasons internalism begs the question.

It strikes me that Parfit and Shafer-Landau’s interpretation of Williams is uncharitable. As we saw in section 1.4.3, the argument can be reconstructed in a way that begs no questions. Williams’s explanatory constraint, as paraphrased in (E-W), doesn’t contain a re-statement of reasons internalism. It asserts that it must be psychologically possible for an agent to be motivated by his or her practical reasons. But, contrary to what Parfit and Shafer-Landau suggest, it does not make any specific claims about how an agent can or cannot be so motivated. After affirming (E-W), Williams separately introduces his Humean premises as a specification of psychologically possible motivation. (W1) and (W2) imply the Hume-Williams thesis, (H). From (H) it follows that it’s only psychologically possible to be motivated by the conclusions of procedural practical reasoning—that is, by practical reasoning from one’s existing motivations. And since conclusions about external reasons are delivered by substantive practical reasoning—reasoning that does not begin from one’s existing motivations—they cannot motivate. Therefore, external reasons fail to satisfy (E-W), and Williams concludes that there are no external reasons. Thus, to refute Williams, the externalist cannot point to circular reasoning in his argument. Williams’s argument can be put in a logically valid
form which begs no questions, as I have done in 1.4.3. It follows that Williams must be challenged with respect to the truth of either his Humean premises—(W1) and (W2)—or to his version of the explanatory constraint, (E-W).

Some philosophers who identify themselves as reasons externalists deny an assumption shared by many prominent reasons internalists. This is the assumption that reasons for action come to motivate through some process of practical deliberation—be it procedural or substantive. Suppose an agent fails to be motivated by some practical reason she has. The idea is that this agent should be able to form a motivation to do what she has reason to do through an exercise of practical reasoning. However, some reasons externalists are not willing to grant such a central role to deliberation in motivating rational action. Instead, these externalists emphasize processes of education, conversion, and socialization as the primary mechanisms through which practical reasons take motivational grip.

John McDowell is an example of this kind of externalist. He does not dispute Williams’s point that practical reasons must be capable of explaining agents’ actions by motivating them (McDowell, 1998, p. 98). Thus he seems committed to some sort of explanatory constraint on reasons. More surprisingly, McDowell concurs with Williams in holding that the only type of practical reasoning is procedural. As a result, in trying to explain an agent’s transition “from not being motivated by a putatively external reason to being motivated by it” McDowell agrees that counterfactual internalists and like-minded
externalists are unlikely to succeed in arguing that practical reasons can motivate independently of agents’ existing motivations (McDowell, 1998, p. 98, 100).

How then, can an “external” reason motivate, if not through substantive practical reasoning? McDowell thinks the debate on the motivational efficacy of substantive reasoning is irrelevant to the question of whether there are “external” reasons. According to McDowell, Williams identified the externalist’s explanatory burden in his 1979 paper. Recall that in trying to discern a way in which an external reasons statement could motivate, Williams offers the hypothesis that “the external reasons statement itself will have to be taken as roughly equivalent to, or at least entailing, the claim that if the agent rationally deliberated, then, whatever motivations he originally had, he would come to be motivated to $\Phi$” (Williams, 1979, in SLC, p. 296). McDowell interjects here, inquiring how the agent may come to deliberate rationally in the first place. There is no reason, he says, to think that the transition from non-rational to rational deliberation must be effected by practical reasoning itself; instead, the agent may come to deliberate rationally through some form of “non-rational alteration.” McDowell discusses two processes other than operations of practical reasoning through which an agent might become a rational deliberator. The first process is through “ethical upbringing in a roughly Aristotelian way” (McDowell, 1998, p. 101, my emphasis). McDowell goes on to describe ethical upbringing as

…a process of habituation into suitable modes of behaviour…of a way of seeing things and of a collection of motivational directions or practical concerns...And if the upbringing has gone as it should, we shall want to say that the way of seeing things…involves considering them aright…(Ibid.).
McDowell speaks of the ideally rational deliberator as one who has been “properly brought up” (Ibid.) Since the end result of proper upbringing is a transition to rational practical deliberation, McDowell surmises that a proper upbringing would enable agents to develop an awareness of external reasons that are not rationally related to one’s existing motivations.

The second process of “non-rational alteration” McDowell calls conversion (McDowell, 1998, pp. 101 – 102; 106-107). If an agent has not been properly brought up, he or she can be transformed into a rational deliberator through conversion. McDowell describes it as a shift in “character” or “motivational orientation” that is not effected by practical reasoning (McDowell, 1998, p. 102). McDowell doesn’t have anything definite to say about how conversion operates, but he does cite religious experience as a particular factor which might effect a conversion, and he characterizes conversion more abstractly as

…a schema for explanations of shifts of character…as the idea of an intelligible shift in motivational orientation that is exactly not effected by inducing a person to discover, by practical reasoning controlled by existing motivations, some internal reasons that he did not previously realize he had (Ibid.).

Although McDowell seems to accept that reasons must be able to motivate and assumes that only procedural reasoning is motivationally efficacious, in my analysis he still qualifies as an externalist about reasons. Despite his concessions to Williams, he maintains that practical reasons can exist independently of any particular agent’s motivations. McDowell acknowledges that some people have not been properly brought
up; he also suggests that, for some people, no process of practical reasoning could effect a transition to correctly recognizing all one’s reasons for action. Nevertheless, McDowell points out that “this has no evident tendency to disrupt the natural connection between being that way [i.e., properly brought up] and considering matters aright” (McDowell, 1998, p. 102). Practical reasons, then, are to be appraised not with reference to any particular agent’s existing motivations, as for the actual internalist, nor to any motivationally effective chain of substantive reasoning, as for the counterfactual internalist, but rather to the inclinations and concerns of a *phronimos*. “*Phronimos*” is the name Aristotle gives to the figure of the ideally practically rational agent. In a reply to McDowell’s main essay on the internalism-externalism debate, Williams takes McDowell to be envisaging the idea of a “correct” deliberator, of an agent who consistently and reliably “considers matters aright,” as Aristotle’s *phronimos* (cf. Williams, 1995b, p. 189). Reasons for action can be, in McDowell’s view, “external” because they are the reasons of a *phronimos*. In this way, McDowell suggests that reasons for action are constituted by how a reference agent would deliberate and decide under given circumstances. Such reasons are reasons for all agents, whether or not any particular person can actually manage to become a *phronimos*.

McDowell is not very clear about how he conceives of the explanatory constraint on reasons. An important question is whether he concurs with Williams’s (E-W), or is instead committed to a claim more along the lines of (E-P)? On the one hand, McDowell never contests Williams’s assumption that if practical reasons exist, then it must be
possible for them to explain action. And he doesn’t deny that to say it is possible for reasons to explain action is to say it is possible for them to motivate action (McDowell, 1998, p. 98). Thus, at the very least, McDowell holds that it must be psychologically possible for at least some people to become a *phronimos* through proper upbringing or conversion. But even this minimal commitment does not decide between (E-W) and (E-P). For it does not settle the question of whether, for McDowell, the possibility of becoming a *phronimos* should be indexed to a *particular agent*, or to a *class* of relevantly similar agents? If the former, McDowell is best read as a reasons internalist. But if the latter, it would be more accurate to class him as an externalist. In his reply to McDowell, Williams (1995b, pp. 191 – 194) argues that only the former alternative is coherent. He also denies McDowell’s charge that Williams’s version of reasons internalism does not allow for rational motivation to result from “non-rational” alteration.\(^{22}\) Williams thus concludes that McDowell’s view ultimately reduces to a sort of Aristotelian reasons internalism.

However, the account of moral properties that McDowell develops in other work suggests that he in fact embraces something closer to (E-P) (cf. McDowell, 1985, in SLC).\(^ {23}\) McDowell holds that through proper upbringing or conversion, human beings can be brought to recognize reasons for action. Furthermore, facts about reasons for

---

\(^{22}\) Indeed, recall that Williams does say explicitly in “Internal and External Reasons” that “there are no fixed boundaries on the continuum from rational thought to inspiration and *conversion*” (Williams, 1979, in SLC, p. 297, my emphasis).

\(^{23}\) David Wong makes this observation in *Natural Moralities* (2006, ch. 3, pp. 196-197).
action are construed by McDowell as facts about moral properties like *being just* or *being wicked* (Miller, 2003, pp. 258-259). McDowell hopes to give a realist account of moral properties that does not posit the existence of spooky supernatural entities, such as Platonic forms. He does so by drawing an analogy between moral properties and secondary qualities (e.g., color). Secondary qualities are dispositions to evoke certain kinds of responses in creatures that have appropriate sensibilities. The secondary quality of *being red*, for instance, is a disposition to evoke experiences of redness in organisms with appropriate perceptual sensibilities. Similarly, McDowell argues, the moral property of *being wicked* is a disposition to merit responses such as resentment or disapproval in human beings capable of moral judgment. In addition, McDowell not only thinks that moral properties elicit moral judgments in human beings, but also that moral judgments necessarily motivate people to act on those judgments. This idea is McDowell’s most interesting contribution to the theory of practical rationality. The effect of this claim, if true, is to undermine the distinction between “internal” and “external” reasons. Practical reasons exist as dispositions to evoke judgments and motivation in human agents. On the other hand, the existence of a practical reason does not depend on its motivating any particular human agent. Thus, for McDowell, reasons for action are “internal” to the motivations of agents with the appropriate sensibilities, but “external” to the motivations of agents who lack the requisite sensibilities.

McDowell’s account of moral properties, in my view, collapses under a deluge of criticism, but the important idea is that McDowell does not think that moral properties
elicit judgments and motivations in all human beings. Instead, they only move those who are “suitably sensitive” in virtue of having been “brought up well.” Moral properties, and therefore reasons for action, consistently and reliably move phronimoi. At the same time, McDowell’s account suggests that human beings generally do have the sensibilities necessary to be affected by moral properties. Thus any human being in which those sensibilities are present and properly trained can “perceive” and by moved by them. Since, on McDowell’s view, moral properties are identical to reasons for action, his account of them implies the following necessary relation between practical reasons and motivation: If $S$ is a human being, and $S$ has a practical reason $R$ to $A$, then it is possible that some human being $S'$ who was properly brought up is motivated to $A$ because of $R$. This thesis fits the Parfit’s explanatory constraint, as stated in (E-P).

A final, radical position is I suspect a worst-case scenario for reasons externalists: reject the explanatory constraint on practical reasons in any form. Parfit entertains this thought in a way that few other externalists do (cf. Parfit & Broome, 1997, pp. 110 – 111). He begins by observing that the explanatory constraint, as Williams articulates it, seems to require the possibility that a normative practical reason can be identical to some explanatory reason for action. Thus if I have a normative reason to take some medicine, then it must be possible for that reason to motivate me to take my medicine. In that case, it must be possible for my normative reason to take my medicine to be the explanatory reason for which I take my medicine. However, Parfit argues that there is an irreducible difference between normative and explanatory reasons: namely, in order to have an
explanatory reason to act, it’s necessary that one must believe that one has this reason before one can be motivated by it. But intuitively it seems that one can have a normative reason to act regardless of whether or not one believes one has it. Williams himself clearly recognizes this distinction in his analysis of the petrol cocktail. If the man who unwittingly mixes petrol with his tonic and drinks the resulting concoction, we can explain what he does by citing an explanatory reason which consists in the fact that he wanted a gin and tonic (Williams, 1979, in SLC, p. 293). On the other hand, this explanatory reason cannot be identical to his best normative reason for action, since intuitively his best normative reason would seem to favor not drinking the petrol cocktail. But if normative and explanatory reasons are fundamentally different, then perhaps normative reasons don’t explain action after all. Moreover, perhaps they don’t need to. The externalist can argue, not implausibly, that the normative force of a reason to protect one’s own health would persist even if nobody could possibly act for this reason. Williams himself has no argument against this brand of externalism, in particular because he never offers an argument in favor of (E-W). This is unfortunate, because I think that much of the resistance to reasons internalism stems from a failure to take seriously the force of this central lemma in Williams’s argument. As a corrective, I shall devote chapter 2 to an articulation and defense of (E-W).

1.6 Conclusion

This completes our review of the internalism-externalism debate about practical rationality. We’ve learned that a crucial issue in this debate is the question of whether it
must be possible for an agent to be motivated by a normative practical reason that he or she is thought to have. Whereas internalists hold that normative reasons must be able to explain action, externalists either deny this claim or qualify it. Another contentious question concerns the proper interpretation of the explanatory constraint. Should it be “possible” for practical reasons to explain in the sense of psychological possibility, such that it’s compatible with the laws of psychology for agents to be motivated by their reasons? And if so, does the explanatory constraint require that if an agent has a practical reason, then it must be possible for him or her to be motivated by that reason? In the next chapter, I deliver a positive answer to both questions. I shall argue that Williams’s version of the explanatory constraint on reasons is indeed warranted. Finally, we should assess the controversy over Williams’s theory of motivation (or WTM), and ask if “external” reasons, which are derived from substantive operations of practical reasoning, can fulfill Williams’s explanatory constraint. If they can, then there must be a way for operations of substantive practical reasoning to motivate agents independently of their existing motivations. In chapter 3, I argue that the empirical evidence for such motivationally effective substantive deliberation is weak. Instead, I contend that the most parsimonious and powerful reasons-based explanations of action, both human and animal, are those which cite instrumental reasons for action as motivating causes. Furthermore, instrumental reasons for action figure in the most powerful explanations of moral action. And once we appreciate this, it can be seen that instrumental considerations
also *justify* action that conforms to moral rules. This last point will be substantiated at length in chapters 4 and 5.
2. Chapter 2: Practical reasons and motivation

2.0 Normative versus explanatory reasons for action

In chapter 1 (section 1.4.3), I gave a reconstruction of Williams’s classic argument for reasons internalism. The second premise of that argument expresses Williams’s explanatory constraint on reasons, or (E-W):

\[
\text{(E-W)} \quad \text{If any agent } S \text{ has a practical reason } R \text{ to } A, \text{ then (as a necessary condition) it is possible that } S \text{ does } A \text{ because of } R, \text{ where it is possible that } S \text{ does } A \text{ because of } R \text{ just in case it is psychologically possible for } R \text{ to motivate } S \text{'s } A\text{-ing.}
\]

Williams’s explanatory constraint tells us that a necessary condition for an agent to have a normative practical reason is that it be psychologically possible for the agent to be motivated by that reason. To say that it’s “psychologically possible” for an agent to be motivated is to say it’s compatible with the laws of psychology for the agent to be motivated. (E-W) implies that normative reasons are capable of explaining action. But in chapter 1, we also saw that Derek Parfit accuses Williams of equivocating between explanatory (or motivating) reasons and normative (or justifying) reasons for action (Parfit & Broome, 1997, p. 113). Explanatory reasons are factors which explain action, but normative reasons are considerations which favor or justify it. Parfit contends that no normative reason can be identical to any explanatory reason, because one’s having an explanatory reason to act requires believing that one has such a reason. And intuitively it
seems that one can have a normative reason to act whether or not one believes it. To repeat Parfit’s example, if I need some medicine to protect my health, then whether or not I believe this I have a normative reason to take my medicine. Second, even if it’s false that I need some medicine to protect my health, but I believe this anyway, I could still have an explanatory reason (but not a normative reason) which could explain why I take my medicine (Parfit & Broome, 1997, p. 113). Parfit summarizes the distinction between normative and explanatory reasons as follows (Ibid.):

…while [explanatory] reasons require that we have some belief, whether or not this belief is true, normative reasons are provided by some truth, whether or not we believe it.

I think this argument of Parfit’s commits a straw man fallacy. But I find it worth introducing this chapter with, because it invites an opportunity to clarify the relationship between normative and explanatory reasons that defenders of (E-W) are committed to. Philosophers who accept (E-W) can appreciate the difference between normative and explanatory reasons as Parfit describes it. Even Williams agrees that people can have normative reasons independently of believing they have them. In his thought experiment involving the man who is unwittingly about to drink a glass full of petrol, Williams grants that the man has a normative reason not to drink its contents even though he does not believe this (cf. Williams, 1979, in SLC, p. 293). This normative reason derives from the fact that there is a “sound deliberative route” from the man’s existing motivations, which presumably includes a motivation not to poison himself, to the conclusion that he should not drink what’s in the glass.
Furthermore, the difference between explanatory and normative reasons that Parfit identifies at best establishes that they are not identical kinds. But Parfit does not argue against the possibility that something could be both an explanatory and a normative reason, even while there are some normative reasons which are not explanatory reasons and some explanatory reasons which are not normative reasons. (E-W) states merely that if an agent has a normative reason to $A$, then it must be possible for that normative reason to contribute to the explanation of his $A$-ing. If the normative reason figures in the explanation of the agent’s $A$-ing, then by definition it is an explanatory reason. In short, (E-W) says that it must be possible for a normative reason to be an explanatory reason. But none of this requires explanatory and normative reasons to be identical kinds, such that any normative reason is an explanatory reason and any explanatory reason is a normative reason.\(^1\) What is required for (E-W) to be true is that all an agent’s normative reasons must be potential explanatory reasons for that agent’s actions, where the relevant meaning of “potentiality” is psychological possibility.

Although this particular argument of Parfit’s fails, it is part of an important challenge to Williams’s explanatory constraint. Parfit suggests that theories of practical rationality that endorse (E-W) mistakenly equate the normative force or “oughtness” of reasons for actions with some kind of motivating force (Parfit & Broome, 1997, pp. 111 – 112, 126 – 128). (E-W) does indeed make it a condition for the truth of normative reasons

\(^1\) Consider an analogy: If $S$ is a human being, then it must be possible for $S$ to be a human being and an animal. This is certainly true, but it doesn’t imply that any human is an animal and any animal is a human!
statements of the form “$S$ has reason $R$ to $A$” that $R$ potentially motivates $S$ to $A$. But Parfit notes that the normative force of such reasons claims seems to persist despite having no effect on agents’ motivations. For instance, if Isabelle needs her medicine to protect her health, it seems to follow that she has a normative reason to take her medicine. Now suppose that Isabelle doesn’t care about her health, and she has no motivation whatsoever to take her medicine. Also, suppose that it’s impossible for Isabelle to acquire a motivation to take her medicine. Even so, it’s tempting to maintain that Isabelle still has a reason to take her medicine. Here we have an intuitively plausible case in which normative force persists in the absence of motivation. If this intuition is well-founded, then, contrary to what (E-W) asserts, an agent can have a normative practical reason despite the impossibility of its motivating the agent.

In what follows, I offer a defense of Williams’s explanatory constraint. I will be particularly concerned to rebut similar versions of the challenge raised by Parfit. The basic point of this challenge can be paraphrased by the slogan that normativity persists in the absence of motivation. But the precise meaning of the challenge to (E-W) is as follows: Agent $S$ can have a normative reason $R$ to do $A$ despite its being psychologically impossible for $R$ to motivate $S$’s $A$-ing. Before I can begin to dismantle Parfit’s challenge, I’ll have to develop a positive argument in support of (E-W). This will be the task of sections 2.1 – 2.4. In section 2.1 I present the argument itself, showing how it unifies various arguments and fragments of arguments scattered throughout Williams’s writings on practical rationality. In section 2.2, I argue that any normative reasons claim of the
form “S has a normative reason to do A” entails that S can do A, where this means that S has both the ability and the opportunity to do A. In section 2.3, I specify the minimal requirements for rational agency, which I understand as the ability to act in a manner favored by normative reasons. I argue that rational agency requires the ability to act for—i.e., because of—normative reasons. Specifically, I contend that a necessary condition for rational agency is the capacity for one’s action to be explained reason-governed practical inference. In section 2.4, I argue that it is possible for an agent to act for normative reasons only if it is psychologically possible that normative reasons motivate the agent to take the actions favored by those reasons. The arguments laid out in sections 2.2 – 2.4 bring me to the conclusion that (E-W) is true. Then, in section 2.5, I respond to various formulations of the main challenge to (E-W)—namely that the normativity of reasons persists in the absence of motivation.

Before going further I should note that whenever I use the term “reason” in isolation, I refer to a normative practical reason. I use the phrases “explanatory reason” or “motivating reason” to designate factors that explain action, rather than justify it.

2.1 The master argument for Williams’s explanatory constraint

So far as I know, Williams has no official argument for his explanatory constraint. However, there is one suggestive passage in “Internal Reasons and the Obscurity of Blame” (Williams, 1995a) where he gestures toward a grounding principle. In this passage, Williams makes a series of assertions about when it is appropriate to blame people or offer them advice (Williams, 1995a, p. 40):
Focussed blame operates in the mode of ‘ought to have’, which has a famous necessary connection with ‘could have’. Focussed blame will go by the board if ‘could have’ is absent. If ‘ought to have’ is appropriate afterwards in the modality of blame, then (roughly) ‘ought to’ was appropriate at the time in the modality of advice. Now ‘ought to’ in the modality of advice implies ‘can’, because advice aims to offer something as a candidate for a deliberative conclusion. If Φ-ing is not available to the agent, ‘You ought to Φ’ cannot function as a piece of advice about what he should now do; when it is a matter of what I am to do, manifestly ‘I cannot’ acts as a stopper.

Williams coins the term “focused blame” to refer to expressions of blame which are applied specifically to the acts or omissions of agents, as opposed to an explanation-oriented blame which we apply, for instance, when we blame a valve for the failure of a rocket. Focused blame may be applied in response to moral offenses, but this is not always the case. It can also apply to non-moral failures, such as mistakes, or missed opportunities. As an example of a non-moral failure, Williams describes a bank robber who is blamed by his accomplice for making an idiotic mistake that results in their arrest. The accomplice need not invoke the system of morality to blame his inept partner. Suppose, for instance, that the robber forgot to keep the bank manager from tripping the silent alarm. Errors like this may stem from carelessness, but not immorality.

Focused blame, then, can be directed at normative failures of a more general sort than immorality. Williams adds that focused blame is not deployed for the sole purpose of condemning normative failures committed in the past. It can also serve a future-oriented capacity, as what he calls a proleptic mechanism that prompts the person being blamed to take a different attitude toward a past action for which he is blamed (Williams, 1995a, pp. 41-2). When focused blame is applied in this capacity, its intended effect is to
get the one being blamed to change his mind about the reasons for which he acted in the past. But the purpose is not to get the recipient of blame to simply recognize in hindsight some past failure in his own practical deliberations. Indeed, the one being blamed may not have committed any such failure at all; at the time, he might have even acted on his best reasons. Rather, what focused blame aims to do is to make the recipient include the blame itself as a consideration in his reassessment of his past actions. When focused blame succeeds in this aim, it may be hoped that the recipient comes to the conclusion that his best reasons for action were not the ones he had acted on after all. Consistent with his procedural conception of practical reasoning, Williams suggests that focused blame can bring about such a process of reassessment by appealing to the recipient’s motivation to have the respect of other people. The more the recipient of blame cares about having the respect of those who blame him, the more likely he is to change his mind about what his past reasons were under the circumstances. In effect, Williams seems to be offering a plausible account of the pragmatic function of blame. We use blame not only to point out the past failures of others to act on their best reasons, but also to give them new reasons by which they can evaluate or re-evaluate their conduct.

Williams has similar comments for advice. Whereas focused blame applies to acts committed in the past, advice applies to acts about to be taken. But as Williams sees it, focused blame and advice are both forward-looking in the sense that they are proleptic mechanisms. When offering advice to someone, Williams observes, we sometimes hope that our advice itself is registered as a consideration which influences the future course of
their practical reflections. And if the people we advise are sufficiently motivated to gain our respect, then that advice in itself will constitute a reason for them to follow it.

In Williams’s discussion of blame and advice, we can find the elements of an argument for (E-W). In the suggestive passage quoted above, Williams treats the principle that ‘ought’ implies ‘can’ as a constraint on when blame or advice can be applied. This idea is surely one of the most ubiquitous in the history of philosophy. It can be traced to Pelagius, Augustine, and Hobbes, though Kant’s invocation of it is probably the most familiar (Vranas, 2007, p. 198, n. 3; Kavka 1986, p. 310). There is too little textual evidence in Williams’s writings to construct a formulation of the principle as he might have understood it. But he does offer a few hints. To begin with, both blame and advice are expressed using the locution “ought.” Blame is expressed by means of “S ought to have done A,” and it’s suggested by Williams that this claim would “go by the board” if it turns out that S could not have done A after all. Likewise, advice is expressed by means of “S ought to A,” and Williams tells us that if A is “not available to the agent,” then “S ought to A” cannot perform the function of advice. More telling are Williams’s observations about the relationship between normative practical reasons, blame, and advice. Focused blame, Williams writes, “involves treating the person who is blamed like someone who had a reason to do the right thing but did not do it” (Williams, 1995a, p. 42, emphasis added). In other words, the function of focused blame is to identify a practical reason that therecipient ought to have acted upon at some past moment, but didn’t (although, as suggested above, he may not have actually had this reason). And if someone
could be blamed for not acting on some practical reason in the past, then that person could have been advised to act on this reason at the time of the action (Ibid.). In effect, focused blame and advice are said to be in the business of identifying reasons for action. A piece of blame articulated as “S ought to have done A” entails that S had a reason to A, and a piece of advice expressed as “S ought to A” entails that S has a reason to A. But Williams also tells us that blame or advice must be retracted whenever the agent “cannot” do A. From these textual snippets I take it that Williams is committed to something like the following thesis, which I offer as a lemma in the argument for Williams’s explanatory constraint:

**Reasons imply ‘can’ (abbreviated “RIC”)**

(RIC) Necessarily, if agent S has a reason to do A, then S can A, where S can A if and only if

(i) S has the ability to A, and

(ii) S has the opportunity to A.

Typically, the “ought” in ‘ought’ implies ‘can’ is understood as the “ought” of moral obligation (as in “S ought to A in virtue of his moral duty to A”). But Williams suggests the principle applies to a broader range of expressions using “ought.” He clearly takes it to apply to advice and focused blame, which he ties to the “ought” of practical rationality (as in “S ought to act, or ought to have acted, on his best practical reason”).

113
Therefore it will be more precise to introduce a cousin of the ‘ought’ implies ‘can’ principle that asserts a necessary connection between a normative practical reason to A, on the one hand, and the ability-plus-opportunity to A, on the other. Thus we have the thesis that “reasons imply can,” or (RIC). My formulation of this principle draws on Peter Vranas’s excellent recent (2007) defense of ‘ought’ implies ‘can’, where (RIC) is in fact a premise in Vranas’s argument for the ‘ought’ implies ‘can’ principle. I also follow Vranas in his definitions of the terms “ability,” “opportunity,” and “implies” (Vranas, 2007, pp. 169 – 170). For an agent to have the ability to do something, she must have the requisite skills, physical capacities, and know-how to do it. And for an agent to have the opportunity to do something, she must be in circumstances which allow her to exercise her ability to do it. In addition, to say that one statement “implies” another is to say that, as a matter of conceptual necessity, the latter statement is true if the former is true.

Though much more could be said about (RIC), the most important task for my purposes is to elaborate a specification of “ability” that supports Williams’s explanatory constraint. Williams offers a starting point when he speaks of people who are unmoved by the proleptic mechanisms served by focused blame and advice. He writes:

Of course, there are some hard cases, people who lie beyond any such mechanism; and it is a support for an account on these lines, that it is precisely people who are regarded as lacking any general disposition to respect the reactions of others that we cease to blame, and regard as hopeless or dangerous characters rather than thinking that blame is appropriate to them. This represents
the absence from their [motivational sets] of anything that can be reached by these mechanisms (Williams, 1995a, p. 43, emphasis added).²

Accordingly, focused blame succeeds as a proleptic mechanism only if the one being blamed has a disposition to gain the respect of others in his motivational set, or \{M\}. Of course this disposition, like any other, may vary in its power to motivate from person to person. When someone who has committed some normative failure is unmoved by the blame of others because he has little or no inclination to gain their respect, Williams claims that it’s *inappropriate* to blame them. This raises an important question: *On what grounds* does Williams think it improper to blame those who are indifferent to the blame of others? Within the framework of Williams’ actual reasons internalism, the most straightforward answer is that an agent unmoved by blame would have no reason to avoid being blamed, because there would be no “sound deliberative route” from the agent’s existing motivational set to the conclusion that he shouldn’t have done what the critic blames him for doing. But from this it does not immediately follow that such a recalcitrant agent cannot be *appropriately* blamed for what he does. That conclusion could be reached, though, if Williams had accepted a thesis which I call *blame internalism*. Blame internalism claims that there is a necessary connection between having a reason to do something and being blameworthy for failing to do it. Thus, if I’m blameworthy for allowing my parents to languish in illness when they reach old age, I must have had a reason not to allow this. Blame internalism appears often in the

² Williams’s shorthand notation for the motivational set of an agent is the letter ‘\(S\).’ But since I use ‘\(S\)’ to designate an agent, I use the symbol ‘\(\{M\}\)’ to stand for the agent’s motivational set.
philosophical literature on practical rationality, and it’s quite plausible on its face (see especially Shafer-Landau, 2003, in SLC, p. 316). For how could anyone be blamed for doing something they never had a reason to do in the first place? Ultimately, to justify William’s claim that it’s inappropriate to blame people for actions they had no reason to avoid doing, we can mount the following argument:

**The impropriety of blame**

(B1)  *Blame internalism*: If S is blameworthy for doing A, then S has (or had) a reason not to do A.

(B2)  *Williams’s reasons internalism (paraphrased)*: If S has (or had) a reason R to A (or not to A), then S has some motivation M to A (or not to A).

(B3)  Therefore, from (B1) and (B2), if S is blameworthy for doing A, then S has some motivation M not to A.

(B4)  Therefore, from (B3) it follows that if S lacks M, then S is not blameworthy for doing A.

(RIC) doesn’t appear in the argument above. But I think Williams must have had some reason to gesture towards it before proceeding to a discussion of blame in his “…Obscurity of Blame” essay. I take it that (RIC) plays a role in the justification of Williams’s reasons internalism, or premise (B2) above. (A more precise formulation of Williams’s position is stated as thesis (W) in chapter 1, section 1.4.3.) It does so, I
believe, by serving as a premise in an argument for Williams’s explanatory constraint (E-W), which is itself a premise in the argument for his reasons internalism (see chapter 1, section 1.4.3). Accordingly, I propose the following “master argument” from (RIC) to Williams’s explanatory constraint:

The “Master Argument” for Williams’s Explanatory Constraint (E-W)

(E1) Reasons imply ‘can’ (paraphrased): If S has a reason R to A, then S has both the ability and the opportunity to A.

(E2) If S has a reason R to A, then S has the ability to A only if it is possible that S does A because of R.

(E3) It is possible that S acts because of R only if it is psychologically possible that R motivates S to A.

Therefore,

(E-W) If S has a reason R to A, then it is psychologically possible that R motivates S to A (this is Williams’s explanatory constraint paraphrased).

In presenting this argument for (E-W) I am no longer engaged in Williams interpretation. The argument represents my own attempt to develop a foundation for a crucial thesis in Williams’s thought. I shall now substantiate each of the premises.
### 2.2 “Reasons Imply Can”

For whatever my intuitions are worth, (RIC) strikes me as nothing less than a conceptual truth. How could someone have a reason to do something she has neither the ability nor the opportunity to do? Some philosophers, however, have argued that it is coherent to suppose that someone “ought” to do that which she cannot do. In his (1971) critique of the ‘ought’ implies ‘can’ principle, Michael Stocker illustrates the case of a man who judges that he ought to pick up broken glass on a path frequented by barefoot children. Suppose, however, that the children tie him securely to a tree, depriving him of the ability to pick up the glass. Stocker thinks it would be perfectly coherent for the man to say that he is being prevented from doing what he ought to do.

In this context it makes no difference whether we understand what this man “ought” to do in the sense of moral obligation, or in the sense of practical rationality. There is no doubt a very close connection between these two understandings of the term, though some philosophers think the relation is more intimate than others. For instance, *moral rationalists* hold it to be a conceptual truth that any moral duty entails a normative practical reason to fulfill the duty. Counterfactual internalists and externalists about reasons are uniformly moral rationalists. Actual reasons internalists have resources for showing how moral conduct can be practically justified, though they usually reject the view that one necessarily has reason to be moral. I have nothing to say about this particular debate in the present chapter (though I’ll have plenty to say in chapter 3).

When I discuss thought experiments originally intended to suggest in favor of or against
the ‘ought’ implies ‘can’ principle, I simply re-interpret the relevant “ought” claims in terms of practical rationality with the understanding that, given the proper background assumptions, sense can be made of these re-interpretations under of any of the major theories of practical rationality discussed in chapter 1.

So what shall the defender of (RIC) say of Stocker’s man tied to a tree, who insists that he “ought” to pick up the broken glass despite not being able to do it? Here is the challenge: if (RIC) were true, it would not be the case that the man has a reason to pick up the glass, since he cannot do it. And yet it wouldn’t seem so odd for the man to complain that he is being prevented from doing what he ought to do. But to this objection the defender of (RIC) can reply that there are multiple senses of “ought,” and while some of them do seem to imply ability-plus-opportunity, some of them do not. Stocker himself describes the ideal-expressing or axiological mode of “ought,” which is understood to mean that it would be “good” or “better” if something were done whether or not it could be done (Stocker, 1971, p. 304). For instance, it would be good thing indeed if I were to develop a cure for every disease. Because I lack the requisite competence, skills, resources, and time, I have neither the ability nor the opportunity to do so. Nevertheless, my lack of ability and opportunity doesn’t negate the thought that it would be a good thing to do if I could do it. In a similar vein, the man who thinks he ought to pick up the glass despite being tied to a tree can be interpreted as expressing “ought” in the ideal-expressing mode. We can understand him to be saying that it would be better for him to
pick up the glass, if he could. (RIC) would not negate an “ought” claim of this type even when the agent cannot do what he ought to do.

Stocker thinks this rebuttal only begs the question because it does not justify the choice of interpreting the man as using “ought” in the ideal-expressing sense, rather than in the sense of moral obligation (Stocker, 1971, p. 312). However, I think the ideal-expressing interpretation is favored by independent considerations which discredit the idea that people sometimes have reason to do what they cannot do. Vranas (2007, p. 186) formulates a potential argument in favor of the view that reasons do not imply ‘can’:

(reasons do not imply ‘can’ (Vranas, 2007))

(1) For any agent S, if it would be good for S to do A, then S has a reason to A.
(2) Some agents cannot do the things it would be good for them to do.
(3) Therefore, some agents cannot do the things they have a reason to do.

The first premise of the argument is flawed. Again, I take it that it would be a very good thing for me to single-handedly cure all diseases. It would also be good for me to transport myself instantaneously and put a stop to every case of child abuse occurring in the world at this very moment. I should also think it a very good thing for me to see into the future and predict every harm that will befall any person I’ve ever cared about. However good these actions sound, and however strongly I might wish to undertake them, what seems obvious is that I have no reason to do any of them. And (RIC) provides
a perfectly plausible explanation: I have no reason to do these things, because I lack the ability to do any of them.

Additional support for (RIC) surfaces with the observation that one’s set of reasons seems to expand and contract in accordance with one’s abilities and/or opportunities. To my ears, it would sound false if the man tied to the tree insisted not just that he ought to pick up the glass, but also that he had a reason to do it. This intuition of mine can be corroborated by imagining cases where it’s perhaps more obvious that the agent cannot do the relevant action. No one can change the past, no matter how much we might wish we could. A woman who was just diagnosed with an advanced stage of breast cancer might regret not getting regular mammograms earlier on in her life. Suppose she acquires the ability to re-live any past moment in her life by snapping her fingers. With this magical new ability, it seems to me that the space of the woman’s reasons would change dramatically. A whole range of new considerations would enter into her practical deliberations, and salient among them might be a reason to go back and get more regular cancer screening. On the other hand, without special time-travel powers, the woman is just like the rest of us. In that case she may sensibly say in hindsight that she had a reason, which only now she fully appreciates, to go in for more regular mammograms, but it would surely sound anomalous for her to say that she now has a reason to change the past. To offer a somewhat more earthly example, suppose I tell you I have a reason to appoint Judge X as the next US Supreme Court justice. Already you might think that’s a very strange thing to say, but in response you might ask politely if I plan to run for
President of the United States. I say no: I have no interest in public office. But still I insist I have a reason to appoint \( X \) as the next justice. At this moment you might think I could benefit from a civics lesson: you inform me that only the President is able to appoint justices of the Supreme Court. But I reply that I already know this, and that even so I have a reason to appoint \( X \) for Supreme Court justice. Here I think you would feel forced, quite justifiably, to conclude that I’m just talking nonsense. By contrast, consider the President of the United States, sitting in the Oval Office. A vacancy in the Supreme Court has just opened up, and the President, being a long-time admirer of Judge \( X \), decides he has a reason to appoint \( X \) to the bench. It seems quite plausible that the President could have this reason, and quite bizarre to suggest that I could. The difference seems to be rooted in the differences between my abilities and those of the President.

(RIC) offers a powerful account of these intuitions: in every case, the presence or absence of a reason for action is a function of the agent’s abilities and/or opportunities. In all the thought experiments just presented, the absence of a requisite ability or opportunity to do \( A \) does seem to imply the absence of a reason to \( A \). At the same time, it seems there are many good things that we wish could be done, but which we don’t think confer reasons for action precisely because we cannot do them. Thus, contrary to Stocker’s judgment, it is appropriate to recognize an axiological sense of “ought” which describes an ideal situation without entailing any normative reasons. And whenever someone judges that they ought to be doing something that cannot do, it is more
consistent with the totality of the considerations just presented to think of him as using “ought” in the ideal-expressing mode.

2.3 Acting for reasons

I now turn to a defense of premise (E2) in the argument for Williams’s explanatory constraint. The premise states that if S has a reason R to do A, then an agent has the ability to A only if it is possible for her to do A because of R. To understand the claim being made by (E2), it’s helpful to recall from chapter 1 (section 1.0) the definition of a rational agent as one whose actions are responsive to normative reasons. In light of this definition, we may ask the following question: given that an agent has reasons for action, in virtue of what does the agent have the ability to be responsive to reasons? (E2) answers this question by making a claim about what constitutes the ability to act rationally. The precise meaning of the clause in the antecedent of (E2), which says that “S has a reason R to do A” is that an agent, S, has a normative reason R which favors or justifies action A. The consequent of (E2) is a claim about what is necessary for an agent to have the ability to respond to reason R. This necessary condition is the possibility that S does A because of reason R. By the phrase, “S does A because of R” I mean that S’s doing A is explained by R. A far less cumbersome way of expressing the idea that “S does A because of R” which I adopt is simply that S does A for R. Therefore, (E2) effectively says this: in order for an agent to be able to respond to a normative reason, it must be possible for the agent to perform that action for that reason. In short, rational agency requires a capacity to act for reasons.
Some creatures are non-rational because they’re incapable of responding to reasons. One consideration which supports (E2) is that the capacity to act for reasons seems to be what distinguishes rational agents from non-rational beings. For instance, (with a few exceptions) human beings are generally rational agents, whereas plants are not. Whether non-human animals are rational agents is a more controversial matter, but many have argued that at least some animal species are capable of rational agency (see in particular Hurley & Nudds, 2006; Bermúdez, 2003). What, then, constitutes the ability to be responsive to reasons? Part of the answer is a capacity for thought. In a recent article, Fred Dretske (2006) attempts to identify the minimal conditions for rational agency. He designates these conditions by the term minimal rationality. According to Dretske, minimal rationality has to do with the way the behavior of an organism is generated. By his definition, minimal rationality “requires that thought be involved in the process by means of which the behavior is produced” (Dretske, 2006, p. 107). The kind of behavior-governing thoughts that Dretske has in mind are intentional states. Intentional states are mental states which “contain” representations of states of affairs. The paradigmatic examples of intentional states are beliefs and desires. Intentional states are said to have representational “contents.” For example, my belief that it’s raining “contains” the proposition that it is raining, and this proposition represents a state of the world as it is or might be. In addition, my desire to attend the Rolling Stones concert “contains” the proposition that I attend the Rolling Stones concert, and this proposition represents a state of the world as I want it to be. The “contents” of intentional states have intentional
properties: the content of a belief may be true or false, and the content of a desire may be fulfilled or unfulfilled. Since intentional states are mental states, creatures that have intentional states must have a brain and nervous system.

According to Dretske, what distinguishes (minimally) rational agents from non-rational creatures is that their behavior is explained by the representational contents of intentional states. He draws an important distinction between a behavior which is explained by an event that represents something, and one which is explained by a representation (Dretske, 2006, pp. 110 - 111). Although the behavior of non-rational creatures can sometimes be explained by features that represent things, their behavior is not explained by representations. For instance, plants lack intentional states, but Dretske acknowledges that some of them are endowed with organic structures capable of representing the world. He gives the example of the scarlet gilia, a plant that changes color from red to white every year in the middle of July. It does so in order to attract different species of pollinators at different times during the flowering season. Earlier in the season, hummingbirds are the main pollinators, and they are more attracted to red blossoms. Later in the season the hummingbirds migrate, and hawkmoths become the primary pollinators. The moths prefer whiter blossoms. The mechanism which brings about the gilia’s change in color is a chemical “switch” in the plant activated around the time when the hummingbirds migrate. Thus, according to Dretske, this chemical switch
represents the arrival of July. When a scarlet gilia changes color, this color change is caused by the chemical switch. But, Dretske argues, the color change is not explained by the fact that in a particular gilia the switch represents a time of year. Instead, the gilia’s color change is caused by its switch, and the switch has this causal role because natural selection favored ancestors of this gilia endowed with similar switches. Since the gilia’s color change is not explained by the representational content of its chemical switch, its behavior is not minimally rational.

By contrast, Dretske contends that the behavior of minimally rational creatures is explained by the representational contents of their intentional states. He supports this point with the case of a foraging bird that eats a poisonous monarch butterfly, and then becomes sick. Later the bird encounters a viceroy butterfly, which is not poisonous but adapted to have similar markings to the monarch. Unable to tell the difference between the two kinds of bugs, the bird avoids the viceroy. In Dretske’s analysis, the best explanation for the bird’s avoidance behavior is that, having learned that a similar-looking bug once made it sick, the bird forms a belief whose representational content is a noxious monarch butterfly. Accordingly, the bird qualifies as a minimally rational agent.

---

3 Of course, the chemical switch can *misrepresent*. It can be activated too soon by unusually hot weather and drought. In that case, the gilia will display white blossoms before the hummingbirds leave and the hawkmoths arrive (Dretske, 2006, p. 112). But this does not make the gilia’s switch any less of a representation. *Any* representation is in principle capable of misrepresenting. Intentional states are no exception. Beliefs “contain” representations, and yet beliefs can be false because their representations sometimes don’t represent accurately. Indeed, some philosophers take the possibility of misrepresenting to be essential for anything to be a representation at all (cf. Neander, 2004).
The appeal of Dretske’s notion of minimal rationality is that it does seem to offer a plausible account of the difference between rational and non-rational creatures. However, Dretske is not concerned to give an account of what it takes for a being to be responsive to *normative reasons*. Dretske sometimes defines minimal rationality in terms of reasons, but he seems to speak of “reasons” in the explanatory rather than the normative sense. For instance, he writes:

> Minimal rationality requires that what is done be done for reasons, but it doesn’t require that it be done for good reasons. Nor does it require reasoning. Although the behavior must be explained by a thought in order to qualify as minimally rational, it needn’t be rationalized or rationally justified by the thought that explains it, and the agent needn’t have computed (reasoned) his way to that result (Dretske, 2006, p. 108).

Since, according to Dretske, the “reasons” which explain the behavior of minimally rational creatures need not justify the behavior, they need not be normative reasons. Instead, the “reasons” Dretske discusses seem strictly to be the representational contents of intentional states. Such representational contents are “reasons” only insofar as they can explain behavior.

Nevertheless, I think Dretske’s account of minimal rationality sheds light on what it takes for an agent to be responsive to *normative* reasons. Because plants lack intentionality, they are not minimally rational in Dretske’s purely explanatory sense. But neither do they respond to normative reasons. Indeed, plants are not the sort of beings capable of having normative reasons for action at all. The scarlet gilia’s color-changing behavior is not of the sort which can be justified from the normative perspective of practical rationality. This is not to suggest that the gilia’s color change cannot be
subjected to normative evaluation at all. Ruth Millikan has famously argued that some features of organisms have what she calls proper functions. The notion of proper functioning is inherently a normative one (Millikan, 1984). If some property $P$ of an organism has some effect $F$ which explains why that organism’s ancestors were favored by natural selection, then $F$ is the proper function of $P$ (Okrent, 2007, pp. 93 – 94). The proper function of the gilia’s chemical switch is to change the plant’s color, because the gilia’s ancestors had similar switches, and the effect of these switches in changing the color of those ancestors enhanced their reproductive success. According to Millikan, an item with a proper function can be evaluated with respect to its success or failure in performing its proper function. Thus we can say that the gilia’s chemical switch ought to make the plant turn from red to white with the arrival of July, because this is the proper function of the switch. I have no quarrel with Millikan on this point, but it should be emphasized that the norm according to which the behavior of the scarlet gilia is evaluated is not a norm of practical rationality. Principles of practical rationality favor actions that either advance some supposedly objective value, such as the welfare of others, or bear some ideal relation to the agent’s motivations, such as maximally satisfying one’s desires. But when a scarlet gilia changes color too soon, as sometimes can happen in unseasonably hot weather, we evaluate its behavior from the normative perspective of proper functioning, and not from the perspective of practical rationality. We do not say that the gilia fails to respond to reasons, but instead that it fails to do that-which-explains-
why-it-exists. Because nothing in the behavior of the gilia or any plant seems to be evaluable by the lights of practical rationality, plants are non-rational creatures.

Dretske tells us that the scarlet gilia is not minimally rational because its behavior is not explained by intentional states. Likewise, I think that plants are not responsive to normative reasons in part because their behavior is not intentional. In my view, intentional behavior (i.e., behavior caused by intentional states) is susceptible to evaluation by the norms of practical rationality. This is because the ability to be responsive to normative reasons requires that the agent’s behavior be explained by a process of practical inference which is governed by normative reasons. Intentional behavior can be explained in this way, since intentional states are capable of bearing inferential relations to one another.

Practical inferences guided by normative reasons can be carried out through intentional states. As a case in point, consider one fascinating experiment conducted with a New Caledonian crow named Betty (Kacelnick, 2006, pp. 101 - 102). In preliminary trials, Betty had been provided with two wires, one straight and the other hooked. With the hooked wire she was able to lift an out-of-reach basket of food sitting at the bottom of a plastic well. Then, in test trials, Betty was given the straight wire only. After repeated attempts, Betty could eventually use several different techniques to bend the wire into a hook, and then retrieve the food basket. Biologist Alex Kacelnick describes just one of Betty’s wire-bending techniques (Kacelnick, 2006, p. 102):

After failing to lift the basket with the straight wire, she took it to a fracture in a nearby plastic tray, wedged the tip there and pulled perpendicularly from the
proximal side, bending the wire until it formed a hook. She then returned to the well, retrieved the basket and ate the food.

It’s very tempting to explain Betty’s responses to the problem she faced in terms of a process of practical inference. An alternative hypothesis might be that Betty’s tool-making in these experiments was the product of an automatic behavioral routine which evolved as a biological adaptation typical of her species. This hypothesis likens Betty’s behavior to that of the Sphex wasp (Dennett, 1984; Carruthers, 2004). The wasp stings a cricket, paralyzing it, and then places it in a burrow with its eggs to provide a food source for its hatchlings. When it captures a cricket, it drags the prey to the entrance of the burrow. It then leaves the cricket at the entrance and goes inside the burrow, apparently to check for intruders. But when an experimenter moves the cricket away from the burrow’s entrance while the wasp is inside, the wasp will repeat the same sequence of behaviors: drag the cricket to the entrance, check inside the burrow, and so on. By repeatedly moving the cricket back, an experimenter can make the wasp repeat this sequence an indefinite number of times. The apparently rigid and automatic nature of the Sphex wasp’s behavior suggests against its being generated by a process of inference. Instead, the behavior might be explained as an innate behavioral program which evolved through natural selection operating on past generations of wasps.4

4 Consequently, the wasp’s behavior cannot be appraised from the perspective of practical rationality. At best, it can be appraised with respect to the proper functioning of its innate routine. That is, it makes sense to say that the wasp ought check its burrow before moving in the cricket, but only because burrow-checking is part of a routine that promotes the wasp’s reproductive success (and promoted the reproductive success of its ancestors). The wasp should check its burrow not because it’s the rational thing to do, but because it’s the adaptive thing to do. By the same token, it makes sense to say that a Sphex wasp which
However, Kacelnick argues that Betty’s behavior is much too flexible and context-specific to flow from an innate routine. Although New Caledonian crows are noted for being “consummate tool-makers and users,” in their natural environments the birds “normally do not have access to pliable material that can be bent into a shape and preserve it” (Kacelnick, 2006, p. 102). Therefore, the more plausible hypothesis is the one which explains Betty’s behavior as a product of thinking (Ibid.). Such thinking could be characterized as a practical inference. On this hypothesis, Betty would have formed a belief that bending the wire would help her get to the food. This belief, when paired with a desire for food, issues in a practical inference which causes Betty to carry out the means to her desired end. A capacity for practical inference such as this would more adequately explain Betty’s tool-making behavior, since the ability to form and act on means-ends beliefs would enable Betty to respond flexibly to novel circumstances.

Suppose this hypothesis holds up, and Betty is indeed thinking her way to action. There is also a strong impression that Betty’s thinking is justified from the perspective of practical rationality. The Instrumentalist Dictum (cf. chapter 1, section 1.1) could quite naturally be invoked to evaluate Betty’s tool-making. Her behavior is favored by this principle because it promotes the satisfaction of Betty’s desires. So, Betty’s practical inferences are reason-governed in the sense that they accord with some principle of

fails to check its burrow is not behaving as it ought, but not because this behavior isn’t rational. Rather, the wasp that doesn’t check its burrow isn’t doing what it ought to do because, in its natural environment, this behavior is maladaptive.
practical rationality. Furthermore, Betty is responsive to normative reasons insofar as her behavior is explained by her reason-governed practical inferences. This conclusion draws support not only from the experiments with Betty, but also from equally impressive experiments with dolphins and other bird species.  

The weight of these observations leads me to affirm (E2). Whereas Betty seems to be a rational agent whose behavior is responsive to normative practical reasons, the scarlet gilia is clearly not. The best explanation for this difference is that Betty, unlike the gilia, can carry out and be moved by reason-governed practical deliberations. But to say that an agent’s action is explained by reasons-governed practical deliberations is equivalent to saying that the agent acts for reasons. Therefore, if an agent has the ability to respond to reasons, as any rational agent must, then that agent acts for reasons.

Betty has the ability to act for reasons in virtue of having intentional states like beliefs and desires. For beliefs and desires are capable of bearing inferential relations to one another in operations of practical reasoning. However, this is not to suggest that intentionality is necessary for the ability to draw inferences. Arguably, a simple calculator can draw inferences even without having any intentional states. My claim is rather that reason-governed practical deliberations do occur through intentional states, and because of this, intentional behavior is evaluable from the perspective of practical rationality.

---

5 For a comprehensive review of philosophical and empirical investigations into the rationality of non-human animals, see the introduction to Hurley and Nudds (2006), and subsequent chapters by Kacelnik, Dretske, Millikan, Bermúdez, Hurley, Clayton et al., Call, Shettleworth & Sutton, and Herman.
Against this argument it might be suggested that there is some other capacity, besides reason-governed thinking and action, which makes a creature responsive to reasons. Consider another gem discussed by Dretske: the blinking reflex. When you poke at my face, I will blink. Obviously I have a reason to prevent you from poking me in the eye. And blinking is certainly an effective means of keeping your finger out of there. Of course, blinking is an involuntary reflex: I don’t have to think to blink. But if that’s true, then I don’t blink because thinking in a reason-governed way causes me to do it. Indeed, it seems incorrect to say that I blink because I have a normative reason to. Blinking is just something I automatically do whenever an object shoots toward my eye. Could blinking then constitute a counterexample to (E2)? In other words, could the blinking reflex exhibit an ability to respond to reasons, yet not be a behavior which is explained by reasons?

No. For normative reasons do not favor blinking. That is, blinking is not the sort of behavior that can be justified by principles of practical rationality, because the blinking reflex is not responsive to reasons in the first place. If I knew that you had no intention to stick your finger in my eye, I wouldn’t have a reason to shut my eye. But even then I would still blink, or at least find it hard not to. It might be true that when you do try to poke me in the eye, I will blink. But in so doing, I am not responding to my reason to keep your finger out of my eye. Rather I blink because, given certain kinds of stimulation, I can’t help myself. Similar comments apply more generally to other reflexive behaviors: I have no more reason to blink than I have to sneeze or startle at a
loud noise. These behaviors, even when performed by thinking agents, do not support an ability to respond to reasons. In fact, my intuition is that the missing element which prevents these behaviors from being *rational* is that they cannot be explained by thinking, understood as reason-governed practical inference. But if this intuition is on target, it only provides additional support to (E2).

### 2.4 Reasons as causes

Premise (E3) of the argument for Williams’s explanatory constraint asserts that the possibility that an agent $S$ can act for (i.e., because of) a reason $R$ requires the *psychological* possibility that $S$ can be motivated by $R$. To say that a behavior is *psychologically possible* is to say that it’s permitted by the laws of human psychology. Further, (E3) states that reasons for action must be capable of motivating an agent. Motivation is assumed here to be a causal notion. In general, I understand a motivation to be a cause of action which is “internal” to the psychology of the agent. An earthquake may be said in elliptical fashion to be the cause of my taking cover, but it is not my *motivation* for taking cover. Strictly speaking, my motivation to take cover was that I wanted to avoid harm. To be sure, the earthquake figures in the explanation of my action, but only insofar as it generated within my psychological constitution a motivation to take cover. (E3) is therefore the claim that if reasons can explain action, then they must be capable of *causing* the agent to take the actions which they favor.

---

6 It will be noted below that some philosophers, particularly Jonathan Dancy, reject this assumption.
A particularly interesting externalist challenge to (E3) comes from Jonathan Dancy (cf. Dancy, 1995, 2000). Dancy holds that an agent’s action can be explained by normative reasons without being caused by them. He thinks that reasons explain action *non-causally*, with the psychological states of an agent playing no role in the explanation. Dancy’s position is distinctive, since he accepts a very general form of an explanatory constraint on normative practical reasons. This general version of the explanatory constraint was presented in chapter 1, section 1.4.3, as thesis (E):

(E) If any agent $S$ has a reason $R$ to $A$, then it is possible that $S$ does $A$ *because of* $R$.

Thus Dancy acknowledges that normative practical reasons must be capable of explaining action. And unlike Parfit, he thinks it must be possible for normative practical reasons to be identical to explanatory reasons (Dancy, 1995, p. 13). What Dancy denies is that *reasons explanations* are *causal* explanations (Dancy, 1995, p. 17). By *reasons explanations*, I mean explanations of action in which normative reasons are cited as the *explanans*. Dancy’s basic thesis is that normative practical reasons are *facts*, or states of the world (Dancy, 1995, p. 13; Dancy, 2000, in SLC, 2007, p. 272). But facts, Dancy acknowledges, are causally impotent, and accordingly they cannot motivate (Dancy, 1995, p. 14). Even so, Dancy insists his view does not fall afoul of (E).

---

7 Dancy is also committed to the converse thesis: that it must be possible for a motivating or explanatory reason to be identical with a normative practical reason (Dancy, 2000, in SLC, p. 268).
Dancy takes issue with an assumption which he believes has gone uncontested in the philosophical theory of motivation. This assumption, which Dancy dubs *psychologism*, is the doctrine that “motivating reasons are psychological states of the agent” (Dancy, 2000, in SLC, 2007, p. 268). The “psychological states” Dancy refers to in particular are intentional states. Thus, psychologism is the thesis that actions are explained by intentional states. Psychologism is held by reasons internalists and externalists alike. What Dancy finds especially objectionable about psychologism is the picture its draws of the relationship between normative and explanatory reasons for action. On this picture, explanatory reasons are instantiated by intentional states, and normative reasons are represented within the contents of intentional states. Normative reasons explain action by virtue of generating the intentional states which explain action (Dancy, 2000, in SLC, pp. 272 – 274). Dancy calls this picture a *three-part story*, which posits that normative reasons explain action through the intermediary of explanatory reasons (Dancy, 2000, in SLC, pp. 267 – 272). If we let the symbol “→” represent an explanatory relation, the three-part story amounts to the following model of reasons explanation: *normative reason* → *explanatory reason* → *action*. According to the three-part story, action is explained by explanatory reasons, which are instantiated in intentional states. Normative reasons both explain explanatory reasons and justify the action which the explanatory reasons explain. Thus, normative reasons explain action by the transitivity of explanation—that is, if A explains B, and B explains C, then A explains
C. We’ll soon see that the three-part model is common to a variety of accounts of how normative reasons figure in the explanation of action.

The three-part story is not specific to either internalist or externalist views of normative reasons. In fact, some version of the three-part story can be found in both internalist and externalist theories. For instance, instrumentalism is an actual internalist theory which subscribes to the three-part model. To see this connection, let’s return to Dretske’s foraging bird. Suppose again that the bird sees a butterfly, but this time it correctly believes the bug it sees to be a poisonous monarch. Thus, the representational content of the bird’s belief is a monarch. Also, the bird has a desire to avoid getting sick. The representational content of this desire is a state of affairs in which the bird is not sick. Given this belief-desire pair, the bird forms the means-ends belief that by avoiding the butterfly, it will avoid getting sick. According to the Instrumentalist Dictum, the bird has a normative reason to avoid the bug it sees, provided that (1) its means-ends belief is justified and (2) by avoiding the monarch it maximizes the satisfaction of its desires. Thus, on an instrumentalist conception of normative practical reasons, the bird’s reason for action in this situation must be individuated with reference to the contents of its beliefs and desires. If the contents of the bird’s beliefs and desires were different, then it would have different normative reasons. (For instance, if the bird had desired some food, correctly believed that the bug it sees is a harmless cricket, and justifiably believed that

---

8 See chapter 1 for a survey of the distinctions between instrumentalism, reasons internalism, and externalism.
eating a cricket would maximize the satisfaction of its desires, it would have had a reason to eat the bug.) Additionally, if the bird avoids the monarch, its normative reason can contribute to the explanation of the bird’s action. For the normative reason figures in the content of the bird’s intentional states, and instrumentalists assume that intentional states—namely, combinations of beliefs and desires—can explain action.⁹

The three-part story is not unique to instrumentalism. Counterfactual internalists and reasons externalists also subscribe to it. Thomas Nagel, for instance, advances a counterfactual internalist view which holds that normative beliefs about objective values are sufficient to motivate action (Nagel, 1970). On Nagel’s view, objective values (e.g., one’s own health, another person’s happiness, one’s duty to care for one’s family, etc.) constitute normative reasons for action. These values figure in the contents of normative beliefs. In turn, normative beliefs generate a motivation in the agent to promote objective values. Nagel’s theory subscribes to the three-part story because normative reasons explain action through the intermediary of intentional states and their contents. Also, Margaret Little (1997) develops an externalist theory of acting for reasons heavily indebted to John McDowell. On Little’s account, normative reasons explain action via intentional states which have both belief-like and desire-like features. These “hybrid”

---

⁹To elaborate, instrumentalists take it for granted that the combination of a desire to E and a belief that A-ing is the best means to E is sufficient to explain an agent’s A-ing. This belief-desire combination has motivational force in virtue of the fact that E appears in the contents of both the desire and the means-ends belief. But having a normative reason to A, on the instrumentalist view, also requires that E appear in both the contents of the desire and the means-ends belief. So, on the instrumentalist account, if an agent satisfies the conditions for having a normative reason, he will also satisfy the conditions for having an explanatory reason which can explain the agent’s taking the action favored by his or her normative reason.
intentional states are like beliefs in that they represent the world as it is, and they are like desires in the sense that they represent a state of the world that the agent is motivated to bring about. The hybrid states that Little speaks of have reasons for action as their representational contents. And according to Little, a virtuous agent will be motivated by such hybrid states to act as reasons require.

Dancy rejects psychologism and the three-part story. He instead holds that normative reasons can explain action directly, without the intermediate presence of intentional states. In other words, Dancy thinks the second “part” in the three-part story should be eliminated. For example: on Dancy’s view, my taking cover during an earthquake is explained neither by my belief that an earthquake has struck, nor my desire not to be harmed. Instead, I take cover because of the fact that I have a reason to avoid harm. Similarly, if I have a reason to save a baby from a burning building, then my saving the baby is explained not by my desire to save him nor by my belief that I have a duty to save him, but by the fact that I have a reason to save him.

One major difficulty with Dancy’s account is that there is solid basis for thinking that normative reasons for action are individuated by agents’ psychological states—particularly their beliefs. This returns us to a point of dispute in the theory of practical rationality broached in chapter 1 (section 1.1). There, it was asked whether an agent can have a reason to act which is based on a false belief. I think the answer must be “yes,”

10 In chapter 3, I shall have more to say about the distinction between beliefs and desires in terms of their opposing “directions of fit.”
because it's evident that actions can sometimes be justified by false beliefs as long as those beliefs are justified. Recall that Richard A. Fumerton argues this point with reference to high-salt diets (cf. chapter 1, section 1.1). Here I reproduce the thought experiment, with minor modifications, in greater detail. Start with the plausible assumption that I have a normative reason to maintain good health. I’m advised by my physician who tells me, on the strength of the most current medical research, that I should maintain a low-salt diet because salt intake is linked to high blood pressure. But suppose that current medical research is systematically mistaken, and contrary to what most medical researchers have concluded, blood pressure is actually a *decreasing* function of salt intake—that is, the more salt in my diet, the lower my blood pressure. Despite these facts it seems that I would still have reason to keep a low-salt diet, precisely because I’m *justified* in the belief that doing so would promote my health. In the absence of reliable alternative sources of evidence, I’d be quite justified in deferring to the authority of medical researchers when I consider what to believe about the health effects of salt consumption, even if it turned out the medical research community was in error. My justified (but false) belief that a low-salt diet will promote my health is a consideration that favors keeping a low-salt diet. Therefore, my belief constitutes a reason for me to avoid too much salt. If I indeed had such a reason, there would seem no way to characterize it independently of what I *believed* and whether or not that belief was justified.
Dancy is aware of this objection. In response, he denies that a false belief can “play the role of a normative reason” (Ibid.). But the way Dancy defends himself against this objection is puzzling. Rather than disputing thought experiments like the one presented in the last paragraph, he actually concedes their point without relinquishing his thesis that reasons for actions are not the representational contents of intentional states.

Dancy writes (Dancy, 2000, in SLC, p. 269):

…one’s having reasonably believed that \( p \) is often offered as a justification of one’s action in cases where it turns out that it was not the case that \( p \). And it may indeed succeed in justifying one’s having acted as one did…without this meaning that it is able to play the role of a normative reason.

So Dancy concedes that a false but reasonably (i.e. justifiably) held belief can justify action, but he does not think it follows from this that such a belief can constitute a normative reason for action. Dancy bolsters this thought with two points. Both are expressed in the following passage (Ibid.):

As I have suggested several times now, my having believed that \( p \) is only able to justify my action if, had it been the case that \( p \), I would have been acting rightly, doing the right thing. The normative reason here would be that \( p \), not my having believed that \( p \). My believing that \( p \) is not a reason for action in either sense (or of either sort); it is at best a justification for my having acted as I did.

In the first sentence of this passage, Dancy provides an analysis of the way in which a false but reasonably (justifiably) held belief can justify an action. He suggests that it is strictly neither a belief of this sort, nor the content of such a belief, that can justify an action; rather, it is the state of affairs which is represented in the content of the reasonably held belief. In the sentence, it would seem that Dancy treats \( p \) as a state of affairs—his view, after all, is that normative reasons are states of affairs. So, applied to
the thought experiment presented in the last paragraph, Dancy’s analysis would go like this: My reasonably held (but false) belief that low-salt diets reduce blood pressure is able to justify my action in consuming less salt if, had it been the case that low-salt diets reduce blood pressure, I would have been acting rightly. (Recall that in the thought experiment it’s supposed to be false that low-salt diets reduce blood pressure.) So the normative reason would be the state of affairs in which low-salt diets reduce blood pressure, and not the fact that I justifiably believed this state of affairs to be the case.

Dancy’s analysis does not capture what is doing the real work in justifying an action carried out on the basis of a false, but reasonably acquired belief. Beliefs represent states of affairs. Dancy suggests above that a state of affairs, reasonably but incorrectly believed, plays the relevant justificatory role. But it seems to me that the justifiability of an action necessarily varies, among other things, with the justifiability of the belief which informs the action. To substantiate this point, I offer the following argument from unpredictable consequences. We commonly assess our reasons for action based on the expected consequences of our actions. But oftentimes the consequences of our actions are simply unpredictable. For instance, we sometimes hear about people dying tragically in freak accidents. A freeway bridge collapses on an SUV carrying a whole family. The lifting apparatus on a garbage truck breaks, letting fly a shard of metal that kills a boy riding in a nearby school bus. On the days these people died, did any of them have reasons to go about their business, given the horrific misfortunes that were about to befall them? On the one hand, there’s a temptation to think their best reasons for action favored
hunkering down in a safe place, canceling the family trip, taking the day off school, and so on. I think this intuition stems from our third-person knowledge about what did happen to the accident victims. In hindsight, we know that on one fateful day, when these doomed individuals carried out their mundane activities, they would involve themselves in a chain of events which would result in their deaths. On the other hand, from the perspective of the accident victims themselves, there was no way that they could reasonably anticipate what was going to happen. No one could have predicted the accidents that killed them. The probability that they would fall victim to such bizarre events was very slim indeed. In effect, the epistemic situation of the accident victims was the same as yours or mine on any given day. We’re well aware of the possibility that something awful might happen to us, but as long as we don’t live in a war zone, we are well-justified in believing that much more likely than not, we’ll be fine. And this justified belief seems sufficient to favor taking the risk of our going about our lives, instead of hiding ourselves away from any and all potential harm.

It seems obvious to me that our reasons for action are determined, in part, by our justified beliefs about the consequences of our actions. This will be so even when our justified beliefs are false. Moreover, the states of affairs which ensue from our actions play no role in justifying what we do in the present, or even what we have done in the past. Suppose we succumb to the intuition that the freak-accident victims had reasons not to go about their plans, given that it would lead to their deaths. If we applied the same criterion in assessing our own reasons for action on any given day, we would be forced
into a kind of agnosticism about our reasons. If a freak accident was to befall me when I leave for work this morning, then my best reason would be to stay home. If not, my best reason would be set about my day as usual. The problem is that I have no way of knowing which chain of events will come to pass. Therefore, at any given moment I have no way of knowing my best reason for action.

However, such agnosticism about my practical reasons is unacceptable. Surely I have some basis for justifying the actions I am about to undertake, even if that justification is based on false premises. On any given day I have reasons to go about my responsibilities and plans, and these reasons consist, in part, of my justified belief that the risk of being the victim of a freak accident is low enough to make it worthwhile for me to go out and get things done. Of course, my belief may be false. It’s possible that one some day, something horrible that no one could ever anticipate might happen to me. But if there is no way for me to form a justified belief about the true chain of events that lies ahead, my reasons for action will still be supported by my justified beliefs about what will happen, come what may. The same comments apply to the actual victims of deadly freak accidents. Either they had to be agnostic about their reasons, or their reasons favored going about their day. Since agnosticism was as unacceptable for them as it is for you or I, their best reasons for action were those supported by their justified beliefs about the world, and not by how the world would in fact turn out for them.

Dancy’s second defense against the objection that false but justified beliefs can be normative reasons is represented in the last sentences of the two passages above. Dancy
counsels us to recognize a distinction between normative reasons, one the one hand, and justifying reasons, on the other. He says, for instance, that “it is important to speak of normative reasons rather than, as is common, of justifying reasons” (Ibid.). From the context, Dancy seems to construe justifying reasons as whatever justifies an action. As for normative reasons, he says that “a good reason for acting is a reason that favours acting” (Ibid.). But this distinction is obscure. How could normative and justifying reasons be different? In what sense could something favor an action without justifying it? I fail to see how these questions could be plausibly answered, and Dancy does not answer them in any event. So I’m inclined to reject Dancy’s normative/justifying distinction in favor of a more standard conception of normative reasons. According to the standard conception, reasons are considerations which justify actions (Lenman, 2009). And if a false belief justifies my taking a certain action, then the belief is a normative reason for action according to the standard conception.

As I understand him, Dancy has to invoke the odd distinction between normative and justifying reasons in order to save his thesis that normative reasons are states of affairs, and not the representational contents of intentional states. A false but justified belief that constitutes a normative reason for action would pose a counterexample to Dancy’s view. For instance, the representational content of the belief that low-salt diets reduce blood pressure is the proposition, Low-salt diets reduce blood pressure. Now if this proposition is justifiably believed, it will be a justifying reason for my decision to maintain a low-salt diet. But, according to Dancy, the proposition cannot be a normative
reason, because in Dancy’s account normative reasons must be states of affairs (e.g., the state of affairs in which low-salt diets reduce blood pressure). As Dancy expresses the point,

…intuitively it seems to be not so much propositions as states of affairs that are our good [i.e., normative] reasons. It is her being ill that gives me reason to send for the doctor, and this is a state of affairs, something that is part of the world, not a proposition. Those who announce that all good reasons are propositions…seem thereby to lose contact with the realities that call for action from us (Dancy, 2000, in SLC, p. 273)

Dancy argues for this contention by describing two popular accounts of propositions, and then asserting that there is no way to make sense of how, on either account, propositions could have the status of normative reasons. On the first account of propositions, the proposition expressed by the sentence “Low-salt diets reduce blood pressure” is the class of possible worlds in which that sentence is true. On the second account, the proposition expressed by the sentence “Low-salt diets reduce blood pressure” is an abstract object consisting of diets, of having low salt content, blood pressure, and a relationship which relates these objects and properties together in a manner analogous to the parts of the sentence (Dancy, 2000, in SLC, p. 274). Dancy thinks it obvious that propositions, on either construal, cannot be reasons for action. They are, as he puts it, “the wrong sort of beast” (Ibid.):

Reasons for action are things like his self-satisfaction, her distress, yesterday’s bad weather, and the current state of the dollar. They cannot be abstract objects of the sort that propositions are generally supposed to be.

Dancy emphasizes repeatedly that a reason for action is not a belief, but rather what is believed (cf. Dancy, 2000, in SLC, p. 267). At the same time, the argument just
summarized is meant to rule out the possibility that what is believed could be a proposition. Instead, what is believed must be a state of affairs. In my view, this argument of Dancy’s fails to show that propositions believed cannot be reasons. For as the examples with salt diets and unpredictable freak accidents indicate, one can have a reason for action based on a justified false belief. Such reasons cannot be individuated without reference to the intentional states—namely, the justified beliefs—of the agent. And if that’s so, then a proposition is precisely the right sort of beast to be a reason. Given that reasons for action are propositions, it’s not difficult to explain why some reasons can be based on false beliefs. A proposition is the sort of thing that can be true or false. Moreover, it’s the sort of thing that can be justified by evidence, even if false. It’s also the sort of thing which can be “contained” by an intentional state, like a belief. Moreover, if a reason for action is the sort of thing which can enter as a premise or conclusion in a practical argument, then it’s quite natural to characterize a reason as a proposition.

Arguments, after all, are composed of propositions, not states of affairs. A state of affairs isn’t the sort of thing which can true or false, justified or unjustified. Nor is it the sort of thing which can be subject to inference. Thus, the claim that reasons cannot be propositions is not nearly so obvious as Dancy would have it. It’s a woefully inadequate basis for Dancy not to recognize justified false beliefs (more precisely, the propositions contained by those beliefs) as bona fide normative reasons.

An even more pressing difficulty for Dancy’s account involves his contention that normative reasons, conceived as facts or states of affairs, explain action. Again, Dancy
accepts the general explanatory constraint, (E). But it’s hard to see how reasons could explain action unless they had some connection with the intentional states of agents. Dancy himself acknowledges in several places that facts are causally impotent (Dancy, 1995, p. 13, 17). Yet he goes on to insist that “justifying reasons are largely normative facts, and…these facts are capable of motivating” (Dancy, 1994, pp. 13 - 14). So how could reasons “motivate” actions, if not through the beliefs, desires, or other intentional states of agents?

Dancy’s response to this worry is to advance a non-causal account of reasons explanations. But before I review Dancy’s non-causalism in detail, let me clarify a difference between Dancy and myself in our use of the term “motivation.” Dancy speaks of “motivation” non-causally. For him, a motivation is an item which explains action, but does not necessarily cause it. I, however, find it very unusual to conceive of motivation non-causally. At the beginning of this section I defined a motivation as a psychological cause of action. So, for consistency’s sake I will resist Dancy’s terminology and continue to treat motivation as an essentially causal notion. In my terms, then, Dancy offers a non-causal account of reasons explanations according to which reasons do explain action, but they do not motivate (i.e., cause) it.

Dancy (1995) presents one argument for a non-causal account of reasons explanations. The argument apparently rests on two assumptions. The first is that actions are essentially the reasons for which they are done. In other words, for any given action, the reasons for which it is done constitute the action’s identity—they are what makes the
action what it is. On a “brutely causal account” of reasons explanations, however, reasons are merely the *causes* of actions. This line of thought leads to the second assumption: in any cause-effect relationship, the cause and effect must be metaphysically *distinct* entities. But there is no way to conceive of the reasons that favor an action independently of the action itself. So it cannot be that normative reasons are the causes of action. This, at any rate, is how I interpret Dancy’s claim that “there is no possible metaphysical distinction between the action and the reasons for it, for if we subtract the reasons from the action, there is not enough left to be an action at all” (Dancy, 1995, pp. 17 – 18).

Alfred Mele offers a powerful analogy in an effort to reduce Dancy’s argument to absurdity (Mele, 2005, p. 52). Suppose I have a genuine $1 bill in my pocket. This genuine bill was produced by the US Treasury Department. However, if the bill had not been produced by the US Treasury, it would not be a genuine dollar. So there is “no possible metaphysical distinction” between this genuine $1 bill and the US Treasury Department. But by Dancy’s reasoning, it would follow that the US Treasury Department could not have produced the $1 bill. This is counterintuitive. A more plausible analysis suggests that the $1 bill essentially has a *causal history* such that it is originally produced by the US Treasury. Clearly this fact does not prevent the US Treasury from causing the bill. By similar reasoning, it may be essential to any given action that it has a causal history such that it was produced by a particular reason or set of reasons. But that does not make it impossible for a reason to cause the relevant action.
Dancy may reply that the analogy does not support Mele’s point. The $1 bill is not “produced” by the US Treasury in any causal sense of the term “produced.” The US Treasury does not “produce” dollars bills by printing green ink onto slips of paper. Instead, it constitutes dollar bills as dollar bills by determining what counts as a dollar. Specifically, the US Treasury endows dollar bills with their essence by, among other things, stamping slips of paper with the insignia of the Treasury. The relation between the US Treasury and the dollar bill is not a causal one, but what might be called a constitutive one. Dancy could just insist that reasons constitute actions in a similar way by endowing them with their essences, but not by causing them.

However, Mele’s analogy will be vindicated if it’s shown that actions are individuated via their causal etiologies. In other words, what gives an action $A$ its identity is that it is caused in a certain way. Furthermore, if the agent’s intentional states play a role in the causal process which individuates an action, then Dancy’s position cannot stand. These propositions are supported by what Mele dubs a causal approach to the explanation of action. The approach treats human actions as events caused by intentional states—beliefs, desires, intentions—or their neural-physiological realizations (Mele, 2005, p. 51). Furthermore, assuming (as Dancy does) that normative practical reasons explain actions, I contend that the causal approach supports the view that normative...
reasons cannot explain action except by virtue of causing it via the intentional states of the agent.\footnote{Here I should stress that I am only assuming that normative reasons explain action. The argument to come seeks only to establish the conditional conclusion that if normative reasons explain actions, then they must explain actions by virtue of causing them. Skeptics about practical rationality may accept this conclusion, but deny that normative reasons can explain action causally. If that’s true, it would follow that reasons don’t explain action at all. But I do not offer a specific argument against such skeptics here. Instead my concern is to refute the claim that normative reasons can explain actions without explaining them causally. This is the claim to which Dancy is committed.}

The arch-nemesis of the causal approach to action is what Mele calls anticausalist teleologism. On the latter view, explanations of action are never causal explanations. Instead they are so-called teleological explanations which cite the aims, goals, or purposes of human agents (Mele, 2005, p. 38). It turns out that Dancy himself is committed to this anti-causalist teleological program (Dancy, 1995, p. 18). According to Mele, the causal approach is superior to non-causalist views such as Dancy’s, because the latter fail to meet a challenge posed more than four decades ago by Donald Davidson in his classic paper “Actions, Reasons, and Causes” (1963). This Davidsonian challenge raises the following question: Is there any plausible account of reasons for action which does not treat them as figuring in the causation of behavior? Mele thinks non-causalists have failed to justify a “yes” response to this question (Mele, 2005, p. 6, 39). Causalists, on the other hand, have sufficient support for a “no” answer. I should quickly emphasize that Davidson is strictly concerned with explanatory rather than normative reasons for action. He defines explanatory reasons in terms of the intentional states of agents—specifically, their beliefs and desires—and he goes on to argue that such reasons are the
causes of action (see in particular Davidson, 1963, in Martin & MacIntyre, 1994, p. 676). But I think that the Davidsonian challenge raises a parallel problem for the likes of Dancy, who is concerned to argue that normative reasons can explain action without explaining it causally. Additionally, I think the same considerations which support the view that explanatory reasons must explain action in virtue of causing it also support the view that normative reasons must explain action in virtue of being causes.

The causalist position is supported by hypothetical cases in which an agent can have several potential explanatory reasons to perform some action, but seems really to act for only one of them—namely, the one that actually bears a causal relationship to the action in question. Mele has devised a provocative thought experiment involving a woman named Ann and a diabolical neuroscientist who manipulates Ann’s behavior in various ways (Mele, 2005, p. 40). I shall recruit Mele’s thought experiment to my own purposes by formulating several variations on it. But first, I indulge in yet another terminological disclaimer. I wish to suggest that a causal account of reasons explanations is compatible with any theory of practical rationality which subscribes to the three-part story. Earlier I described three different versions of the three-part story, one being associated with instrumentalism, another with counterfactual internalism, and a third with reasons externalism. These versions differ with respect to the type of intentional states said to motivate action. Because I want to preserve the understanding that a causal
account of reasons explanation is neutral between these versions, I will use “intention” as a catch-all term for any type of intentional state.  

Let’s turn to the thought experiment. Suppose Ann gets up to open a window. At the time she gets up, she has the intention to get a better view of the street. Call this intention $O$. But just before she opens the window, Ann acquires an additional intention to let in some fresh air. Call this new intention $N$. Now Ann has two intentions: $O$ and $N$. Suppose, however, that an ingenious neuroscientist makes the neural realization of $N$ incapable of having any effect on Ann’s bodily movements. The neuroscientist does not extinguish $N$ or its neural realization; she only keeps them from having their usual effects on Ann’s behavior. On the other hand, the neuroscientist allows the neural realization of $O$ to cause Ann’s bodily movements as usual. As things stand, then, the neural realization of only one of Ann’s two intentions is causing her movements. Consequently, Anne opens the window. But which intention explains Ann’s opening it? Is it $N$? $O$? Both of them perhaps? Or is there no determinate answer? The intuitive answer, according to Mele, is that $O$ explains Ann’s action whereas $N$ does not. And the natural explanation for this intuition seems to be that only $O$ plays a causal role in bringing about Ann’s action. This supports the causal approach to the explanation of action, because the only intention which seems suited to explaining Ann’s action is the one that causes it. On

---

12 The philosophical literature is rather divided on how intentions should be defined. Some philosophers (e.g. Davidson, 1963) analyze the intention to do $A$ in terms of an agent’s desire to achieve $E$ and belief that by $A$-ing, he will achieve $E$. Others (e.g. Mele, 2005) treat intentions as intentional states which are irreducible to combinations of desires and beliefs. In what follows I do not commit myself to either of these accounts.
these grounds Mele concludes that for any explanation of action, whatever is cited as the \textit{explanans} must \textit{cause} the action. I concur with Mele’s conclusions.

Of course, Dancy’s theory is a non-causal account of how normative \textit{reasons}, not intentions, can explain actions. The thought experiment with Ann merely suggests that intentions explain action in virtue of causing it, e.g., by producing bodily movements. But Dancy may reply that even if intentions explain actions by causing them, it does not follow that \textit{reasons} explain action causally. Even if $O$ is the intention that is causally responsible for Ann’s opening the window, her action may still be explained by \textit{reasons} which do not in any way cause her to act. Let us then embellish Ann’s story and assume that she has more than one normative practical reason to open the window. Suppose that Ann hears what sounds like police sirens blaring outside her window. This gives her a reason to open the window and look outside, since this is not something Ann is used to hearing in her tranquil bourgeois suburb. She could investigate the noise by opening the window to get a better view of the street. Call this reason $R-O$. Meanwhile, assume that Ann once again forms an intention $O$ to open the window so she can get a better view of the street. But just when Ann is about to open the window, it occurs to her that it’s so stuffy in her upstairs room that she’s ready to faint. This, too, gives Ann a reason to open the window. Call this reason $R-N$. At that moment Ann promptly forms an intention $N$ to open the window, so she can let in some fresh air. However, assume once again that a diabolical neuroscientist suppresses the effect that intention $N$ has on Ann’s bodily
movements. On the other hand, the neuroscientist allows $O$’s bodily effects to go through as usual.

Now, let it be that Ann opens the window. What explains this? If Dancy is correct, and (1) normative practical reasons can explain action, (2) such reasons do not cause actions, (3) such reasons are not intentions, and (4) actions need not be explained by intentional states, then we are at liberty to explain Ann’s opening the window by citing either $R-O$ or $R-N$ (or both). Dancy himself offers no principled basis for distinguishing the reasons Ann has and the reasons for which she acts. Without this, any or even all of Ann’s reasons favoring the same action will do to explain that action. However, my intuition in the case at hand is that Ann’s action cannot be explained by reason $R-N$. The best explanation I can muster for this intuition is simply that $R-N$ does not cause Ann’s action. This lends support to the causalist view that normative reasons explain actions only if they cause actions. I do not think my intuition merely reflects a causalist theoretical prejudice; the intuition leads me to the theory, not the other way around.

Even if the above embellishment on Ann’s story is persuasive, it still doesn’t show that reason $R-O$ causes Ann to open the window. All I’ve shown is that $R-O$ might be cited to explain Ann’s action, while Ann’s intention $O$ causes it. Could this be a mere accidental association? I think not: $R-O$ explains Ann’s opening the window by way of some non-accidental connection with intention $O$. If this is correct, then it would have to be the case that if intention $O$ did not cause Ann to open the window, reason $R-O$ would
not explain Ann’s opening the window. To support this point, let’s adjust Ann’s case a little more.\textsuperscript{13} Suppose that Ann has both the intentions $O$ and $N$ to open the window. But this time the diabolical neuroscientist remotely takes control of Ann’s muscles from the time she gets up to open the window to the point when she opens it. At the same time, the neuroscientist makes sure that Ann’s bodily movements are perfectly in line with the way that Ann’s intentions \textit{would} have caused her to move. But now \textit{none} of Ann’s intentions are causing her to move. Instead, Ann’s movements are caused by the neuroscientist’s remote control body-controller. The salient question now is: which of Ann’s supposed reasons for opening the window, $R-O$ or $R-N$, explains her action? My intuition is that the answer must be \textit{neither of them}. And as before, it’s apparent that in this case Ann’s reasons fail to explain her actions precisely because none of her intentions are causing her behavior.

If reasons are normative facts or states of affairs, as Dancy conceives them, it’s difficult to see how they could explain an action if not through the agent’s intentional states. Return to Ann and the body-controlling neuroscientist. Suppose the sound of police sirens echoes through Ann’s sleepy suburb, but the neuroscientist temporarily shuts off Ann’s ears so that she doesn’t hear a thing. Then the neuroscientist again takes control of Ann’s muscles and makes her go to the window. Why did she do that? Again, we can suppose the sound of police sirens gives Ann a reason to get a better view of the

\textsuperscript{13} This adjustment is inspired by Mele’s critique of Scott Sehon’s anti-causalist reply to the Davidsonian challenge in Mele, 2005, pp. 50 – 51.
street. But since Ann did not hear the sirens, she could never form the belief that *sirens are blaring outside*, and consequently she couldn’t have formed the intention to get a better view of the street. This scenario in effect removes any possibility that Ann’s belief could serve as an intermediate factor in the explanation of her action. Since Dancy denies that intentional states are intermediary factors in reasons explanations, it would seem he could allow the following reasons explanation to be true:

(O) The reason Ann goes to the window is that police sirens were blaring outside.

Given the details of the case, I fail to see how (O) could be true. Intuitively, the correct explanation for Ann’s going to the window is that the neuroscientist took control of her body. By hypothesis Ann does have a normative reason to get a better view of the street, but it seems incorrect to cite that reason in explaining her action. The three-part story offers an attractive explanation for why (O) is false. Ann’s going to the window is not explained by her normative reason to get a better view, because the neuroscientist prevented her from forming beliefs in connection with this reason.

I’ve been arguing that if a normative practical reason *R* favors action *A*, then it can explain *A* only if the agent forms an intentional state, *I*, that *causes* her to *A*. Moreover, the relationship between *R* and *I* must be non-accidental; there must be some regularity that connects the intentional states which cause an action to the reasons which favor that action. I submit that this regularity is a causal relationship between practical
reasoning and the formation of intentional states, as I illustrated earlier with the example of Betty the crow. For suppose that no such causal relationship holds. We may do so by means of yet another embellishment on Ann’s situation (see and cf. Mele, 2005, p. 61). The sound of police sirens blaring outside once again gives Ann a reason, $R-O$, to get a better view of the street. But imagine now that the neuroscientist hypnotizes Ann so that whenever she hears police sirens, she acquires an intention $O$ to get a better view of the street by opening the window, and she carries out her intention. Now there is a regular connection between Ann’s reason $R-O$ and her intention $O$. But is it right to say that $R-O$ explains Ann’s opening the window? I don’t think so, because it seems to me that Ann’s hypnotic spell, and not her reason $R-O$, is causally responsible for her intention $O$. This suggests that not just any regular connection between $R-O$ and $O$ makes it true that $R-O$ explains Ann’s action. The connection must be a causal one. Let us then imagine that there was in fact a causal relationship between $R-O$ and $O$. Then the reason explains the action in a straightforward way: if $R-O$ causes $O$, and $O$ causes Ann to open the window, then $R-O$ causes Ann to open the window. In my view, we should cash out this causal relationship as one which operates through a process of reason-governed practical inference: Ann concludes, by means of a reason-governed practical syllogism, that she has a reason $R-O$ to open the window. This operation of reasoning causes her to form an intention $O$ to open the window, and this intention in turn causes Ann to open the window. Ann’s practical reasoning is made possible by inferential relations that hold among her intentional states. For instance, hearing the police sirens outside, Ann forms
the belief that a lot of police are nearby. She then judges that if the police are nearby, something must be wrong. But if something is wrong, her family may be threatened. And if she looked outside the window she might be able to assess whatever threat there may be. And since she has a reason to look out for the safety of her family, she should assess anything that potentially threatens them. And so on. These premises will all be represented in the contents of Ann’s intentional states. From them, she may conclude that she has a reason to look out the window, and form an intention which motivates her to act accordingly.

The foregoing series of thought experiments support premise (E3) of the argument for Williams’s explanatory constraint. In the last variation of the case of “Ann and the neuroscientist,” we see that it’s possible for a normative practical reason to explain an action only if it’s possible for the reason to cause the action via that agent’s intentional states. I cannot here articulate a complete account of what sort of causal relationship holds between reasons for action and intentional states. However, it was argued in section 2.3 that a necessary condition for the ability to act for reasons is the capacity to act in accordance with reasons-governed practical inference. In any case, the totality of the considerations presented suggest in favor of a three-part model of reasons explanations according to which reasons cause intentional states, and intentional states in turn cause action. Furthermore, we may identify the agent’s set of potentially action-causing intentional states as his or her set of motivations, or \{M\}. This brings us to the
conclusion that we’ve been after: If agent $S$ has a reason $R$ to $A$, and if it is possible for $S$ to $A$ because of $R$, then it is possible that $S$ is motivated by $R$ to do $A$.

Dancy, however, offers several more arguments in defense of a non-causal account of reasons explanations, and it would be remiss not to address some of them. In one interesting section, he presents a rebuttal to a claim often adduced as evidence for the three-part model. This is the “claim that, if it is not the case that $p$, the reason why the agent acted cannot be that $p$ and must be that he believed that $p$” (Dancy, 2000, in SLC, p. 276). We can make sense of this claim by thinking back to Dretske’s bird. Upon seeing a viceroy butterfly which it mistakes for a poisonous monarch, the bird avoids the viceroy. What was the bird’s reason for avoiding it? The claim Dancy wishes to discredit applies to this case as follows: since it’s false that the bird sees a monarch, the bird’s reason for avoiding the bug it sees cannot be that it’s a monarch. Instead, the bird’s reason for avoiding the bug must be that it believes falsely that the bug is a monarch. This claim supports that three-part model of reasons explanations. To resist it, Dancy argues that in ordinary discourse people sensibly adduce what he calls non-factive reasons explanations. This type of reasons explanation, Dancy tells us, is nothing less than “a way of explaining an action by laying out the considerations in light of which the agent acted without committing ourselves to things being as the agent there conceived them to be” (Dancy, 2000, in SLC, p. 277). Dancy’s formal definition for a non-factive explanation is as follows (Dancy, 2000, in SLC, p. 277):
Factive vs. Non-factive reasons explanations

A reasons explanation of the form

The reason why it is the case that \( p \) is \( q \)

is *factive* just in case it entails both \( p \) and \( q \). Otherwise, the explanation is non-factive.

To use James Lenman’s example, a factive reasons explanation would be

(C) Caesar died because Brutus plotted his death

since it can be read as implying both that Caesar died and Brutus so plotted (Lenman, 2009). Dancy, on the other hand, wishes something like the following to be an example of a non-factive reasons explanation (cf. Dancy, 2000, in SLC, pp. 277-8):

(P) His reason for doing it was that it would increase his pension

If this reasons explanation were factive, then it would entail that the fellow in question did what he did, and that his doing it would increase his pension. But if that were so, Dancy points out, the following modification would be self-contradictory:
(P') His reason for doing it was that it would increase his pension, but in fact he was quite wrong about that.

If (P) were factive, then (P') would imply a contradiction. For it would imply three things: that the fellow did what he did, and that his doing it would increase his pension, but doing it would not increase his pension. To Dancy’s ears, however, there is no contradiction in (P'). Dancy concludes that (P) must be a non-factive reasons explanation. If that’s the case, then (P) is a reasons explanation that does not, or need not, explain by citing the intentional states of the relevant agent. For even if the fellow’s action would not increase his pension, it does not follow that his reason for doing it cannot be that it would increase his pension. In effect, (P) can be a true reasons explanation for the fellow’s action in spite of the fact that things are not as he supposed. What makes it true? Like any non-factive explanation, (P) explains in virtue of laying out the considerations in light of which the agent acted. So, if I understand Dancy, (P) can be true in virtue of correctly citing the reasons for which the agent took himself to act. This means that (P) cites the same reasons that the fellow himself would cite, if he were asked to explain why he did what he did. That said, (P) should not be equated with the first-person reasons explanation that the fellow would give for his own action. There is a crucial difference: the man’s first-person explanation would be factive, implying among other things that his action really would increase his pension, whereas (P), being non-factive, does not imply this.
Regarding whether there are any non-factive explanations, my semantic intuitions diverge from Dancy’s. Let’s continue to suppose that what the man does would not in fact increase his position. Dancy contends that (P) is non-factive on the grounds that (P’) is not contradictory. However, even if (P’) is not a contradiction, I doubt that it follows from this fact alone that (P) is non-factive. To my ears, (P) is false, and (P’) is not contradictory because the inclusion of “but in fact…” renders the latter true.

Upon being informed that what the man does would not increase his pension, my natural inclination is to regard (P) as false. There is some basis for this inclination. That the man’s action would have increased his pension couldn’t be his reason for doing it, because it’s evident that a non-actual state of affairs cannot explain something which actually happens. Sound does not travel in outer space, but this is not because space is a perfect vacuum. So far as we know, perfect vacuums don’t exist. The correct explanation is that sounds do not travel in outer space because space is close enough to being a perfect vacuum that sound cannot be transmitted through it. The proponent of a three-part model of reasons explanations can accept this point. He would only need to add that an agent’s action can be explained by his or her false belief in a non-actual state of affairs. Dancy’s view, however, seems very much in tension with the idea that the non-actual cannot explain the actual. After all Dancy believes that normative reasons are facts which can explain action without the involvement of beliefs or any other of the agent’s intentional states. This position also raises the vexing metaphysical question of how an agent’s reason for action qua normative fact could be a non-actual state of affairs. Dancy
offers us no assistance in imagining what an answer to this question might be, but then again I doubt any plausible answer is to be found.

My comments arguing for the falsity of (P) suggest that no explanation in which the *explanans* is a non-actual state of affairs can be a correct one. Even more unfortunate for Dancy’s account is the observation that (P) would ring true if it were added that the fellow *falsely believes* that doing it would increase his pension. Given this extra information, I find it most natural to interpret (P) as an elliptical phrasing of a reasons explanation which could be stated precisely as follows:

(P*)  His reason for doing it was that *he believed falsely* that it would increase his pension.

If the man’s action would not increase his pension, I would take both (P) and (P’) to express truths just in case they can be interpreted along the lines of (P*). These semantic temptations deserve to be explained—or explained away. One response available to Dancy is that (P*) may be a true reasons explanation, but the reason it cites is of the wrong sort. The fellow may have an *explanatory* reason, grounded in his false belief, which explains his doing it. But Dancy is concerned to give an account of reasons explanations in which the reason cited as the *explanans* is also a normative reason. Dancy could maintain that no normative reason can be grounded in a false belief. However, I argued earlier that normative reasons can issue from false *but justified* beliefs. If my
earlier argument is cogent, then the fellow can have a normative reason to do what he justifiably believes will increase his pension. And in that case, even when we interpret the fellow’s “reason” to do it in the normative mode, I can maintain my intuition that (P) will be true in virtue of being interpreted along the lines of (P*).

The so-called three-part model for reasons explanations is compatible with the semantic intuitions I report above. Dancy’s non-causal account is not. On the three-part model, beliefs, and perhaps other intentional states, have the power to explain action. Intentional states can be caused by a process of practical inference, and normative reasons for action can govern such a process (if so, the relevant agent will be practically rational). So, reasons cause intentional states, which in turn cause actions. It follows by the transitivity of explanation that normative reasons cause actions.

2.5 Lingering doubts about (E-W)

On the strength of the points I’ve made in favor of (E1), (E2), and (E3), I affirm that the argument for (E-W) is sound. But in spite of my efforts to construct a foundation for (E-W), the reader may still have reservations about it. In particular, there is the objection noted in section 2.0 as the main challenge to (E-W): viz., the claim that the normativity of reasons persists in the absence of motivation. (E-W) tells us that an agent can have a normative practical reason only if it’s psychologically possible for him or her to be motivated by this reason. But philosophers have objected to a version of the ‘ought’ implies ‘can’ principle which similarly claims that one ought to do something only if he or she can be psychologically motivated to do it. These objections can extend to (E-W).
In particular, Michael Stocker offers a possible case where normative intuition suggests that a person ought to do something, despite being psychologically “unable” to do it (Stocker, 1971, p. 311). Consider a man “unable” to bring himself to put his arm in a cesspool to retrieve his wallet. Despite this, it’s not hard to imagine circumstances where it would seem intuitively true that the man has a reason to get his wallet. Peter Vranas cites the same thought experiment to show that an agent’s being psychologically “unable” to do something does not imply that the person cannot do it (Vranas, 2007, p. 169). That is to say, an inability to motivate oneself to do something doesn’t amount to lacking an ability-plus-opportunity to do it. Vranas also mentions how people with addictive disorders (e.g., kleptomania) or substance addictions (e.g., alcoholism, nicotine addiction) are psychologically “unable” to refrain from doing what they ought to do. Although Vranas describes such people as acting in violation of their moral duties, it’s not too much of a stretch to think of them as acting contrary to their best practical reasons (Vranas, 2007, p. 183).

However, I do not think Stocker and Vranas’s examples represent genuine counterexamples to (E-W). It’s compatible with (E-W) to suppose that the people described are acting against their best practical reasons. Defenders of (E-W) can maintain that even if someone is psychologically “unable” to motivate himself to do something, it may still be psychologically possible for him to do it. In ordinary speech, we may say that Stocker’s man is “unable” to put his hand in the cesspool because he has extreme difficulty in overcoming the revulsion he feels at the prospect of submerging his hand in
waste. But it doesn’t necessarily follow that it’s psychologically impossible for the man to do this, because the psychological impossibility of doing something is an “inability” in a much stronger sense. When we want to know whether it’s psychologically possible for someone to have a motivation, the question we should ask is whether there is any way, compatible with the laws of psychology, in which the agent could acquire an intentional state that motivates him to perform the action. Moreover, if we want to know whether it’s psychologically possible for someone to be motivated by a reason, we should ask whether there is any way, compatible with the laws of psychology, that the agent could acquire a motivating intentional state by means of a reason-governed practical inference. The defender of (E-W) can resist Stocker’s suggestion that the man “cannot” act in this stronger sense. Perhaps, with some coaching, the man could find within himself some untapped reservoir of determination, and thereby acquire what motivation he needs to get his wallet. Perhaps the same result could be achieved by a combination of punishments and incentives. Similar comments apply to people blighted by addiction. Vranas notes that nicotine addicts, kleptomaniacs, and others strictly speaking can and regularly do control their impulses. Many nicotine addicts don’t smoke where it’s prohibited, and kleptomaniacs avoid stealing when they know they’re likely to be caught (Vranas, 2007, p. 184). People’s motivations can be primed in many ways, and the psychological impossibility of motivation should not be confused with the mere difficulty of acquiring it.

That said, we can certainly conceive of a kleptomaniac or a nicotine addict who cannot resist his impulses because it’s psychologically impossible for him to do so. Such
characters may even exist. In these far-fetched instances, (E-W) would entail that the reasons which we’re used to ascribing to people do not in fact apply to them. This indeed seems to be a bullet that Williams has to bite, given his suggestion (discussed in section 2.1) that it’s inappropriate to blame people who cannot be motivated by our blame. Yet many of us are ready to insist that these people are appropriate subjects of our blame. In his (2001) postscript to “Internal and External Reasons,” Williams attempts to dull the force of this critique by noting that there are all sorts of normative terms that we may appropriately direct at “recalcitrant or unconvinced agents” (Williams, 2001, p. 95). For example, of a husband who callously refuses to stop mistreating his wife, we may say that he is “inconsiderate, or cruel, or selfish, or imprudent; or that things, and he, would be a lot nicer if he were so motivated” to act otherwise (Williams, 2001, p. 95; Williams, 1979, in SLC, p. 297). In this context, Williams does not interpret such normative terms to be carrying the implication that the man fails to act on reasons he has (e.g., a reason not to mistreat your wife). The problem, Williams tells us, is not that people like the abusive husband are acting contrary to their best reasons. Rather,

…the problem is precisely that that they are not. Let \(N\) stand in for some normative term: if the critic expresses himself by saying “There is a reason for this man to behave differently to these people,” then what he says is of the form “There are considerations about these people’s welfare, interests, and so on such that it is \(N\) that this man should treat those considerations as reasons.” What can we take \(N\) to be? It does not seem to me that there is anything in this way of putting the situation which takes us beyond understanding \(N\) as, very roughly speaking, “better” (Williams, 2001, p. 96).

This passage can be understood in light of what was earlier identified as the *axiological* or “ideal-expressing” meaning of the term “ought.” Return to the Stocker’s man tied to
the tree (section 2.2): he may complain that the children who tied him up are preventing him from doing what he *ought* to do—viz., pick up the glass from under their feet.

Stocker recognizes the possibility of interpreting the man’s “ought” claim as expressing what would be *better* for him to do *if* he could do it, rather than a claim about an obligation he is being prevented from fulfilling in the present situation. In the quoted passage it sounds as though Williams is helping himself to a similar rejoinder. By condemning the man who mistreats his wife as cruel, insensitive, and so forth we do not accuse him of defying his reasons. Instead, we are lamenting the fact that there exist people for whom considerations about the well-being of others are not reasons for action.

We saw that Stocker is unimpressed by such appeals to the ideal-expressing “ought,” and it’s evident that Williams’s opponents would be too. Thomas Scanlon, in particular, points out that Williams neglects an important aspect of what is going on when we condemn the abusive husband cruel, insensitive, etc. We are in effect accusing him …of a kind of deficiency, namely a failure to be moved by certain considerations that we regard as reasons. (What else is it to be inconsiderate, cruel, insensitive, and so on?) If it is a deficiency for the man to fail to see these considerations as reasons, it would seem that they must be reasons for him. (If they are not, how can it be a deficiency for him to fail to recognize them?) (Scanlon, 1998, p. 367).

Scanlon challenges Williams and his defenders to provide plausible ways of conceiving cruelty, insensitivity, etc., other than as failures to see certain considerations as reasons. He also asks for some explanation as to why, in criticizing people like the abusive husband, it seems that we invoke reasons which apply *to him*, regardless of what his motivations are. Surprisingly, Williams concedes very much to this challenge of
Scanlon’s. With respect to the abusive husband, Williams even agrees that “in some sense these reasons apply to him; certainly he cannot head off the criticism by saying that the reasons do not apply to him because he does not have that kind of [motivational set]” (Williams, 2001, p. 96). However, if by calling the abusive husband cruel or insensitive we only mean to say that the world would be better off were he to recognize certain considerations as reasons, it’s difficult to see how these reasons would be reasons which apply to him. But in criticizing those who do not act as we think they should, we seem to be doing much more than this. We seem to be accusing people of failing to recognize certain considerations as reasons which are properly theirs.

Williams has an answer to Scanlon’s challenge, though he himself doesn’t press it. The main idea behind this rebuttal is what Williams calls the obscurity of blame. According to Williams, “it must very often be obscure to those who are blaming, and quite probably indeterminate in itself, what the motivational state is of the person being blamed” (Williams, 1995a, p. 43). When we apply focused blame to someone, we can never be sure whether the one we are blaming will be motivated by our blame to re-evaluate the reasons for which she acted. It will always be unclear whether any particular application of focused blame will succeed in its function as a proleptic mechanism. However, it seems to me that people are sufficiently interested in influencing the behavior of others that they would be more likely to err on the side of blaming those who cannot be moved by blame. Also, I take it to be a fact that most people are motivated to gain the respect of at least some others, which makes it a reasonable expectation that
most people’s motivations can be reached by the blame of someone.\textsuperscript{14} Hence I think there is a well-founded presumption that people are generally susceptible to focused blame. And although the man so callously resolved to keep mistreating his wife may try to thwart criticism by claiming he doesn’t have “that kind” of motivational set, I doubt many people would take him at his word. Rather, I’d expect that people would be prepared to continue blaming him.

The availability of focused blame as a strategy for controlling the behavior of others explains how it can make sense, on William’s view, to fault the abusive man for failing to act on reasons that apply specifically to him. However, it does so in a different way than Williams’s opponents would have it. The point of focused blame is not limited to calling attention to the reasons that an agent failed to act on, or informing him about reasons he didn’t know he had. As I noted in section 2.1, the goal of blame can be to give someone a new reason that he didn’t have before by prompting him to re-evaluate his past conduct. Any putative reason that we urge others to adopt through our blame can be said to be properly theirs. Williams’s view does not deny this, but it does tell us that there is a catch—namely, that people’s specific motivations limit the extent to which we can successfully and appropriately try to instill reasons in others.

Even this appeal to focused blame may not be enough to assuage resistance to (E-W). Williams’s critics might reply that the function of blaming others does not have to be limited to changing their behavior. The argument about the impropriety of blame (section

\textsuperscript{14} I discuss evidence supporting a basic human need for “relatedness” in chapter 5.
2.1), which I attributed to Williams, suggests that people who are truly “beyond the pale,” who cannot in any possible way be motivated to act differently than we think they ought to act, are not blameworthy. Yet some of us might have the powerful gut feeling that even such absolutely recalcitrant persons are worthy targets of our blame.

However, I think this is a point with which the defender of (E-W) can readily agree. For the proponent of (E-W) could think of blame along the lines of David Hume’s account of moral judgment. As noted in section 2.1, one type of blame is a reaction to moral offenses. It is a kind of moral judgment. Hume held the position that the primary aim of moral judgment is the evaluation of character. In Hume’s view, the object of a moral judgment is some character trait of a particular person, such as greed or benevolence (Cohon, 2004; 2006, p. 257; Taylor, 2006, p. 278). Character traits are judged as being virtuous or vicious. The virtues and vices are those character traits that produce sentiments of approval and disapproval, respectively, in whoever contemplates the trait from an impartial or “common point of view” (Cohon, 2006, p. 257). Virtues tend to elicit approval in observers because they are agreeable or useful (advantageous) to either their possessor or to others. Vices, on the other hand, elicit disapproval because they are disagreeable or disadvantageous either to their possessor or to others (Cohon, 2004; Taylor, 2006, p. 279). Feelings of approval or disapproval in response to the perception of virtue or vice are caused by what Hume believes to be a basic human capacity for sympathy. Hume speaks of sympathy as a psychological mechanism that allows human beings to interpret, share in, and respond to one another’s emotions.
Thus we approve of the virtues because we sympathize with those who are benefited by virtuous characters, and we disapprove of vice because we sympathize with those who are disadvantaged by the vicious. Accordingly, Hume thought that people’s moral judgments are made not with regard to their self-interest or from their particular points of view, but instead from a “common point of view, from which they might survey their object, and which might cause it to appear the same to all of them” (Hume, 3.3.1.30, in Norton & Norton, 2000, p. 377). The common point of view is the perspective of those who are directly affected by the action of the person being judged. Thus the virtues are approved not because they are agreeable or advantageous to the person making the moral judgment, but because they tend to the benefit of society as a whole (Cohon, 2004).

This quick summary of Hume’s moral philosophy sheds light on a possible function of blame which Williams overlooks. It is a function which, I suggest, could be separate and independent of the function of identifying an agent’s reasons for action. Blame is a means of evaluating a person’s character and its effects on others. Therefore, on a Humean account of blame, we can blame the abusive husband for the harm he does to his wife. Our blame can also serve as a warning to others to use caution in their dealings with this man, for it can be reasonably expected that his wife will not be the sole victim of his callousness. It might also serve as a call for a collectively imposed punishment, such as ostracism, to protect society from persons as incorrigible as he. In Hume’s ethics, the appropriateness of moral approval or disapproval is not in any way
contingent upon whether the person being judged can be motivated to change his or her behavior. Thus, if we subscribe to Humean ethics, and if reasons for action are derived from an agent’s motivations, then blame can serve the function just described without implying that someone has failed to act on his or her reasons.

The Humean account of blame just sketched enables us to make sense of how recalcitrant agents like the abusive husband might be blamed, without necessarily suggesting that the subjects of blame are acting against their reasons. But to accept this account comes with a cost. It is incompatible with blame internalism—the thesis that, necessarily, if anyone is blameworthy for some action, then he or she must have a reason to act otherwise (cf. section 2.1). Judging from Williams’s comments about the propriety of blame (quoted above in section 2.1), it seems that he takes blame internalism to be unassailable. This puts Williams in a rather awkward position with respect to the abusive husband, who cannot possibly be motivated to treat his wife better. On the one hand, Williams’s allegiance to reasons internalism requires him to hold that the man has no reason to stop mistreating his wife. But on the other hand, if blame internalism is true, then this man would not be blameworthy for mistreating his wife. As I’ve noted, this is a difficult conclusion to accept. Williams, however, does not fully appreciate the advantages of rejecting blame internalism while accepting that hopelessly immoral agents are blameworthy in a Humean sense. This package of positions allows us to blame figures like the abusive husband, while recognizing all the considerations adduced in favor of (E-W). Alternatively, if we concede that the abusive husband can
appropriately be blamed, *and* we hold fast to blame internalism, then we’d have to conclude that the abusive husband has reasons which cannot possibly motivate him. This conclusion, of course, is in tension with (E-W), and to accept it would require us to ignore all that has been said in favor of (E-W). By now, I hope to have convinced the reader that abandoning (E-W) is no less contrary to intuition than denying blame internalism.

### 2.6 Conclusion

For some time now it’s been popular to invoke the distinction between normative and explanatory reasons to undercut Williams’s arguments for reasons internalism. I have tried to show that those who rely on this move have failed to take seriously Williams’s explanatory constraint on reasons. Unfortunately, Williams himself does little to defend the thesis which lies at the very heart of his theory of practical reason. My objective has been to fill this gap in Williams’s thought. The conclusion I come to is that if normative reasons for action exist, then it must be psychologically possible for them to motivate agents. To the objection that normative reasons for action may exist even when they cannot motivate agents, I reply that reasons must have some necessary connection to motivation, for this connection seems essential to the very concept of rational agency. Furthermore a theory which holds normative reasons hostage to the potential motivations of agents need not do violence to normative intuition. Once the truth of (E-W) is conceded, it would have the effect of setting the terms of debate in philosophical discourse about practical rationality. It would limit the viable alternatives to those
accounts that can supply a plausible theory of the relationship between practical reasons and motivation. Accordingly, the best account of practical rationality should be compatible with what is known about the causal mechanisms that govern action and decision-making. This ultimately pushes the debate about the nature of practical reason into the domains of ethology and the social sciences. Philosophers should welcome this development, because it carries the promise of ending old stalemates among rival positions that could not be decided when thinking about practical rationality was confined to the armchair.
3. Chapter 3: Humean moral motivation

3.0 Introduction

In chapter 2 it was argued that reasons for action, if they exist, must be capable of motivating agents. This is the idea expressed by Williams’s explanatory constraint, or (E-W). In this chapter I begin to develop an answer to the “Why be moral?” question. Any such answer should tell us the reasons we have to act morally. On the basis of (E-W), I assume that the only serious candidates for a solution to “Why be moral?” are those which tell a plausible story about how a moral reason can come to motivate people. The thesis I wish to defend here, and in subsequent chapters, is that an instrumentalist theory of reasons provides the most plausible account of how moral reasons motivate. Because of this explanatory advantage, I think we must look to instrumentalism for an answer to “Why be moral?” The goal of the present chapter is to establish a presumption in favor of instrumentalism. My strategy is to begin by showing that competing conceptions of practical rationality inform competing accounts of how people are motivated to act morally. After distinguishing the most prominent philosophical theories of moral motivation, I argue that a Humean theory, which assumes that people are motivated to act morally by instrumental reasons, better satisfies two criteria for the choice of psychological hypotheses.

Section 3.1 is devoted to ground-clearing. My aim there is to clarify the connection between the “Why be moral?” question and the debate about the nature of practical rationality. There is a widespread concern that instrumentalism about practical
reason leads to the skeptical conclusion that reasons to be moral cannot apply to people who just don’t care about doing the right thing. An important attraction of counterfactual internalist and externalist theories of practical reasons is that they allow for moral reasons to apply to people regardless of their cares and inclinations.¹ Section 3.2 reviews the literature on moral motivation, with particular attention to the role that practical reasons are said to play in motivating moral action. Sections 3.3 and 3.4 identify two criteria for theory choice that favor motivational Humeanism over its rivals. Finally, section 3.5 rebuts an important objection to my arguments for the presumptive superiority of a Humean theory of moral motivation.

### 3.1 Moral action and practical reason

The debate about practical rationality carries far-reaching implications for the “Why be moral?” question. There is a perception that procedural accounts of practical rationality support various forms of moral skepticism, including skepticism about whether there is any reason to be moral.² The thinking behind these skepticisms can be captured in rather dramatic fashion by an argument I call the Ted Bundy Paradox. The premises express some commonsensical platitudes about morality, as well as an attractive theory of practical reasons. But if these premises are true, they entail the skeptical thesis

---

¹ I survey the distinctions among instrumentalism, actual reasons internalism, counterfactual reasons internalism, and reasons externalism in chapter 1.
² In chapter 1, section 1.4.1, I give an exposition of the difference between procedural and substantive forms of practical rationality.
that there is no reason to be moral. In this section, I review several alternative rejoinders to the Ted Bundy Paradox (henceforth “TB”). First, though, I present the paradox itself:

The Ted Bundy Paradox (TB)

(1) Moral rationalism: Necessarily, if agent S is morally obligated to do (or refrain from doing) action A, then S’s best reason for action favors doing (or refraining from) A.

(2) Instrumentalist dictum: S’s best reason favors doing A if and only if A-ing maximizes the satisfaction of S’s desires.

(3) Murdering innocent people maximizes the satisfaction of Ted Bundy’s desires.

Therefore:

(4) From 2 and 3, Ted Bundy’s best reason for action favors murdering innocent people.

(5) From 1 and 4, it is not the case that Ted Bundy has a moral obligation to refrain from murdering innocent people.

The first premise of TB is a thesis known among moral philosophers as moral rationalism. It asserts that moral obligations necessarily entail reasons for action. Moral rationalism is quite intuitive on its face: I consider myself to have a moral obligation to care for my parents when they grow old, and it seems obvious that by dint of this
obligation alone that I have a \textit{reason} to care for my parents. How could it be otherwise? To some philosophers, it sounds utterly contradictory to say that I have a moral obligation to care for my parents, \textit{despite} having no reason to do it. Moral rationalism is appealing because it rules out the possibility of disconnect between reasons and obligations. It tells us that \textit{necessarily}, if someone has an obligation to care for his or her parents, then he or she has a reason to do it and that’s the end of it.

The second premise of the Bundy Paradox is familiar from chapter 1 (section 1.1). It’s a paraphrase of the Instrumentalist Dictum. Of course, instrumentalism is hardly without its critics. Still, if it is not the most popular view of practical rationality in philosophy, it’s treated by its critics as the default paradigm which has yet to be overturned. Instrumentalism is an appealing view on its face. Its appeal lies in the asserted connection between reasons for action and desire. The fact that I want to get a PhD does seem to constitute a reason for me to get a PhD. And, moreover, if doing something would maximize the aggregate satisfaction of all my wants, then I would seem to have a better reason to do it than anything else.

The third premise is not meant to be a rigorous psychological assessment of the \textit{actual} Ted Bundy—the serial killer who was executed in 1989 for savagely murdering over 30 people. That Bundy derived a kind of sick satisfaction from raping and murdering innocent people is well-known. But there is nothing in particular about Ted Bundy the man that warrants the conclusions of the Ted Bundy Paradox. Rather, Bundy is just a case in point of a widely understood fact: that oftentimes people act on strong
desires to do immoral deeds. His name can easily be substituted for Adolf Hitler, Josef Stalin, or any other wretched individual that comes to mind.

The sub-conclusion in line (4) of the Bundy Paradox gains plausibility when we consider a simple question: Why do bad people do bad things? Presumably because it’s what they most want to do. If murdering innocent people was not what Ted Bundy most wanted to do, then we should be surprised that he did it again and again. But when this suggestion is put together with the Instrumentalist Dictum, we are led down a rather unsettling path. If Bundy most wants to murder, then according to the Instrumentalist Dictum his best reason is to do just that. At that point, we become awash in skepticism. An operation of modus tollens on lines (1) and (4) leads to the conclusion that Ted Bundy does not after all have a moral obligation to refrain from murdering innocent people.

Of course, few of us are willing to submit to these conclusions. The reason is that they run counter to another maxim that commands broad acceptance in moral discourse. I follow Stephen Finley and Mark Schroeder in calling this idea moral absolutism (or MA) (Finlay & Schroeder, 2008, section 1.2.2; see also Joyce, 2001, in SLC, p. 26):

\[
\text{(MA) Moral Absolutism:}
\]

(i) Necessarily, if agent S morally ought to A, then S ought to A regardless of whether A-ing satisfies any of S’s desires or furthers S’s interests.

\[\text{3 In this respect, Ted Bundy is the same figure as Hobbes’s fool or Hume’s sensible knave. For a discussion of these characters in Hobbes and Hume, see Darwall, 1995, pp. 72-3 and 309, and the Introduction to this dissertation.}\]
(ii) Necessarily, if $S$ is blameworthy for doing $A$, then $S$ is blameworthy regardless of whether $A$-ing satisfies any of $S$’s desires or furthers $S$’s interests.

Moral absolutism is the idea that some actions are morally wrong and blameworthy for any agent, no matter what motivations they might have. (MA) tells us that Ted Bundy cannot escape the duty not to murder just because he wants to do it, and this is quite consistent with intuition.

The Ted Bundy Paradox raises a dilemma between conceding to all the argument’s premises, on the one hand, and holding fast to moral absolutism, on the other. Some philosophers have taken the first option. Ethical egoists, for instance, hold that each person only has a moral duty to promote his or her own self-interest (Rachels & Rachels, 2007, p. 69). An ethical egoist can accept all the other premises of the Ted Bundy Paradox, including the Instrumentalist Dictum. Also, Richard Joyce (2001) holds that moral absolutism, if true, implies the existence of external reasons for action which a person can have regardless of his or her motivations. Joyce insists that there can be no such reasons, precisely because they cannot motivate agents to whom they apply (cf. Joyce, 2001, in SLC, pp. 28 – 31).

Philosophers unwilling to give up on moral absolutism reject at least one premise of TB. This amounts to three alternative lines of defense in support of (MA): (1) reject moral rationalism, (2) reject the Instrumentalist Dictum, or (3) reject, or qualify, the

---

assumption that a person can maximize the satisfaction of her desires by way of immoral conduct. Here it’s fruitful to consider the oeuvre of Philippa Foot, who has in fact mounted all three defenses at different points in her career. For instance, in “Morality as a System of Hypothetical Imperatives” (1972) (abbreviated “MSHI”), Foot rejects moral rationalism while not disputing the Instrumentalist Dictum. Foot’s argument in “MSHI” rests on an analogy between moral requirements and the requirements of etiquette. It starts from the observation that the world is full of people who could care less about being polite. As Foot sees it, it sounds odd to say of people who don’t care about etiquette that they fail to be rational. Ordinarily, we apply pejorative adjectives such as “coarse” or “uncouth” to such people, but at least to Foot, it’s inappropriate to call them “irrational.” Based on these observations Foot surmises that the rules of etiquette apply to all persons, but they do not supply reasons for all persons to abide by them. Foot famously suggests that the same is true of moral rules (Foot, 1972, in SLC, pp. 288-89):

The fact is that the man who rejects morality because he sees no reason to obey its rules can be convicted of villainy but not of inconsistency. Nor will his action necessarily be irrational. Irrational actions are those in which a man in some way defeats his own purposes, doing what is calculated to be disadvantageous or to frustrate his ends. Immorality does not necessarily involve any such thing.

Thus, on Foot’s view, we can coherently say of Ted Bundy that he, like everyone else, has a duty not to murder people, even while we concede that he has no reason to do what morality demands of him. Foot is generally interpreted as assuming in “MSHI” some
form of actual reasons internalism, if not instrumentalism.\textsuperscript{5} However, Foot would later repudiate the main thesis of “MSHI,” having been convinced that moral rationalism is true after all (see Foot 1994, 1995, 2001, and 2002).

This brings us to a second possible defense of moral absolutism: one can deny the Instrumentalist Dictum, and indeed any form of actual reasons internalism which makes having a practical reason contingent on the existing motivations of the agent. Instead, one could place bets on some version of counterfactual internalism or reasons externalism, and argue that reasons for action derive from objective values about what is relevantly good or worth achieving (see chapter 1, sections 1.5 – 1.6). One who takes this position could then negate line (4) of the Bundy Paradox, and maintain that even Bundy himself has a reason to refrain from murdering innocent people, regardless of whatever his desires may be. Philippa Foot has recently come to favor this defense of (MA) (see

\begin{footnotesize}
\footnotetext{Foot herself never makes it clear whether she endorses instrumentalism, or another variant of actual reasons internalism which has been called \textit{rational egoism} (Shafer-Landau, in SL, 2007, p. 176). Rational egoism is the view that one has a reason to \textit{A} if and only if \textit{A}-ing will promote one’s self-interest. Rational egoism is a distinct thesis from instrumentalism, because that which promotes one’s “self-interest” may not be equivalent to what promotes the maximal satisfaction of one’s desires. Instrumentalism defines rational action as action which secures a general non-moral end—namely, the satisfaction of desire. Rational egoism may identify rational action as that which achieves a more restricted non-moral end—say, the satisfaction of \textit{self-regarding} desires. For example, a singularly altruistic person may desire to promote the desires of others, but at great cost to other desires she might have for herself. According to this type of rational egoist, acting to satisfy her altruistic desires would run counter to the woman’s (best) reasons for action, even if she maximized the satisfaction of all her desires thereby. This is because \textit{only} self-regarding desires confer practical reasons. Gavin Lawrence attributes this second version of non-moral rational egoism to Foot during the early to middle stages of her career (Lawrence, 1995, p. 105). Alternatively, rational egoism may define the end to be achieved by rational action as a moral one, where what is said to be in a person’s “self-interest” is specified by a substantive moral conception of a person’s well-being. But in that case, rationalism egoism would be a substantive, rather than a procedural, account of practical rationality.}
\end{footnotesize}
especially Foot 1995, 2001). Now she is probably best interpreted as a counterfactual internalist. It would seem that the Recent Foot endorses a substantive account of practical reasoning, since she contends that facts about “goodness” have a necessary connection to practical rationality (Foot, 2001, p. 63). Her internalism comes through in her commitment to the thesis that reasons for action must be capable of motivating agents. With respect to practical reasons for moral action, Recent Foot acquiesces to Hume’s suggestion that morality must be “necessarily practical, serving to produce or prevent action” (Foot, 1995, pp. 3-4).

A third defense of moral absolutism against the Ted Bundy Paradox retains a firm commitment to instrumentalism. We may call this the instrumental defense of morality. Philosophers who mount the instrumental defense target the third premise of TB. They insist that Bundy, or most any person other than Bundy, does not in fact maximize the satisfaction of his or her desires by acting immorally. The instrumental defense has been articulated with impressive sophistication by the philosophical descendants of Hobbes and Hume, including David Gauthier, Gregory Kavka, Kenneth Binmore, and others (see especially Gauthier, 1986; Kavka, 1986, 1995; Binmore 1994, 1998, 2005). Even Foot presented a version of it early on in her career (cf. Lawrence, 1995, especially pp. 106 – 110). The central idea of the instrumental defense is that for most people, most of the

---

6 My recounting of the vicissitudes of Foot’s thought draws on Gavin Lawrence’s 1995 tribute, “The Rationality of Morality.” Lawrence associates a name with each of three major periods in Foot’s development: “Early Foot” refers to the period surrounding the 1958 article “Moral Beliefs,” “Middle Foot” is designated as the author of the classic “Morality as a System of Hypothetical Imperatives” (1972), and finally, “Recent Foot” is the author of all work published by Foot after 1994.
time, it is instrumentally rational to follow moral rules. In other words, conformity with the requirements of morality tends to maximize the satisfaction of most people’s desires. Hobbes and Hume both held that the function of morality is to enable peaceful and cooperative interactions that would advance the mutual benefit of all members of society (cf. Kavka, 1985, in SL, p. 162; Darwall, 1995, pp. 290 – 292). The details of different versions of the instrumental defense vary considerably. Nevertheless, the unifying claim is that moral action is rationalized by the Instrumentalist Dictum. Thus, a possible objection along these lines to premise 3 of the Bundy Paradox would accuse Ted Bundy of being instrumentally irrational: although Bundy believes that he will maximize the satisfaction of his desires by murdering innocent people, his belief is probably false. Perhaps he doesn’t fully appreciate the costs of being caught and punished for his crimes. Or perhaps he doesn’t accurately estimate the probability of being caught in the first place.

However, proponents of the instrumental defense do not hold that moral conduct maximally satisfies the desires of every individual without exception. When discussing Hobbes’s reply to the fool, Kavka does not deny the possibility that a person might come upon a golden opportunity to acquire very large gains by immoral means, and at very low risk of being punished. However, Kavka maintains that such opportunities are in fact very rare, and that typical human biases such as self-deception, the over-valuation of short-term rewards, and the over-estimation of one’s knowledge are likely to make any attempt to find golden opportunities far too error-prone to be worthwhile (Kavka, 1995, pp. 25 –
Though it’s possible that Ted Bundy could indulge his bloodlust without fear of getting caught, the danger of misidentifying an opportunity to kill with impunity might make it instrumentally irrational for Bundy to do violence. Thus, Hobbesians and Humeans endorse some form of rule-consequentialism according to which following moral rules without exception is rationalized by its reliable tendency to maximize the satisfaction of any (or nearly any) individual’s desires. In chapter 5 (section 5.4.1), I will defend such a rule-bound version of the instrumental defense of morality.

Many moral absolutists charge that the instrumental defense of morality does not really address TB. The basic problem raised by Bundy Paradox was that there seems to be no way to simultaneously affirm an instrumentalist theory of reasons, moral rationalism, and moral absolutism. Ultimately, the Hobbesian or Humean must acknowledge the possibility that it may be rational for someone to violate moral requirements. No matter how remote a possibility it is, we can at least imagine a situation in which Ted Bundy justifiably believes that he will maximize the satisfaction of his desires by murdering people. And in that case he will be instrumentally rational. But if moral rationalism is true, then Bundy would not have a duty to refrain from murder, and this contradicts moral absolutism.

Thus, among philosophers who would defend moral absolutism against TB, there are two contending packages of views. The first package consists of moral absolutism and instrumentalism about reasons. Moral absolutists who pursue the instrumental defense of morality must also argue against moral rationalism, if they are to rebut the
challenge articulated in the last paragraph. One way to make this work is to provide a
direct refutation of moral rationalism, as Foot did in “MSHI.” I call this the direct
critique of moral rationalism. A second potential strategy is the one I will pursue in
chapter 5. Rather than showing moral rationalism to be false, one can discredit it by
explaining away its intuitive appeal. I call this the deflationary critique of moral
rationalism. It seems to me that both the direct and deflationary critiques are needed to
successfully persuade people to abandon moral rationalism. A major obstacle to arguing
persuasively against moral rationalism is that so many philosophers uphold it as a
conceptual truth. Again, the suggestion that a person could have a moral obligation, but
no reason to act on that obligation, is simply contradictory to some ears. The aim of the
deflationary critique is to provide a plausible explanation for why, if moral rationalism is
false, so many people consider it to be true. The explanation might run as follows:
suppose it does happen to be the case that most people have reasons to act morally.
Insofar as this idea carries some intuitive weight, it could easily be over-generalized into
the thesis that people necessarily have reasons to act morally. In my view, the
instrumental defense of morality is central to this deflationary critique of moral
rationalism. For it can demonstrate not only that most people have (instrumental) reasons
to be moral, but also why the intuition supporting this claim might be mischaracterized
by moral rationalists as the much stronger claim that moral obligations necessarily impart
reasons.
The second package of views compatible with moral absolutism denies the instrumentalist theory of reasons, but affirms moral rationalism. If moral obligations apply to people regardless of their desires, and obligations necessarily entail practical reasons for the agent obliged, then what is needed is an account of practical rationality which makes sense of the idea that an agent can have reasons to act regardless of what desires he or she has. As already noted, some form of counterfactual internalism or reasons externalism, if true, could pay off this theoretical debt.

In this chapter I argue in favor of the first package of views supporting moral absolutism, and against the second. Much of the effort necessary to criticize the second package has already been done. In chapter 1, I defended Bernard Williams’s argument for actual reasons internalism. The main claim of this argument is that there are no substantive principles of practical reason, because there is no way compatible with the laws of psychology for such principles to motivate people. If Williams’s argument succeeds, then counterfactual internalism will be false. In chapter 2, I developed an argument for Williams’s explanatory constraint, or (E-W). This argument undercuts reasons externalism, which allows for an agent to have a practical reason even if it’s impossible for him or her to be motivated by that reason. Thus, the first two chapters have already chipped away at the moral absolutism-moral rationalism package. Throughout the rest of this chapter, I will keep chipping, and I will begin to adduce considerations in favor of an instrumentalist account of rationality. My basic thesis shall be that if practical reasons for action can motivate agents, as (E-W) requires them to, then
all practical reasons must be instrumental reasons. For instrumentalism is compatible
with the best explanation of how reasons for action can be motivationally effective in the
way that (E-W) requires.

3.2 Moral motivation: Humeanism vs. Anti-Humeanism

3.2.1 Practical rationality, moral motivation, and “Why be moral?”

The debate about the nature of practical rationality and the “Why be moral?”
question converge on the topic of moral motivation. Moral motivation refers to the theory
of the psychological causes that motivate or explain moral action. Moral action refers to
action that complies with the requirements of morality. What is interesting is that
different philosophers’ views on moral motivation vary directly with their opinions about
practical rationality. Of all the various hypotheses that philosophers have adduced to
explain moral motivation, normative practical reasons play some explanatory role. The
usual suggestion is that people act morally because they have a normative reason to do it.
Indeed, reasons internalists who hold both moral rationalism and absolutism must be
committed to this suggestion. For the combination of absolutism and rationalism implies
that all persons necessarily have a reason to be moral. But if so, then whatever the
reasons to be moral are, they cannot be contingent on any particular person’s motivations.
Otherwise, people who happen to lack motivations to act morally (e.g., Ted Bundy)
would also lack reasons to be moral, and this is contrary to the hypothesis of moral
rationalism. At the same time, reasons internalists accept Williams’s explanatory
constraint on reasons, which requires that it be psychologically possible for any
normative reason to motivate the agent who bears the reason. It follows that a reasons internalist committed to both moral rationalism and absolutism must show how there can be reasons to be moral which are not derivative of any particular agent’s motivations, but which can motivate agents nonetheless. To pay these theoretical debts, such philosophers develop accounts of moral motivation in which all persons necessarily have reasons to be moral, regardless of any particular person’s existing motivations, but moral reasons are capable of generating new motivations in (rational) agents to act morally. An exponent of this position, discussed below, is Thomas Nagel.

However, moral absolutists who reject either moral rationalism or reasons internalism also have something to gain by explaining moral motivation in terms of normative reasons. In section 3.1, we saw that instrumentalism appears to be incompatible with the combination of moral rationalism and moral absolutism. (Recall that in chapter 1 (section 1.3) I characterized instrumentalism as a species of reasons internalism.) However, I noted that instrumentalists who are moral absolutists can diminish the appeal of moral rationalism by showing that most people typically do have reason to be moral. Given the instrumentalist’s commitment to the explanatory constraint on reasons, an instrumentalist account of how people can have reason to be moral winds up including an account of how people can be motivated by instrumental reasons that favor moral action.

Finally, even moral absolutists who accept reasons externalism and reject Williams’s explanatory constraint have an interest in saying something about moral
motivation. These philosophers tend to view moral judgment as essential to a capacity to be responsive to practical reasons. However, since externalists reject Williams’s explanatory constraint, they are sometimes challenged by their rivals to explain the tight connection between moral judgment and moral motivation (this is the “basic phenomenon of moral motivation,” discussed below). Consequently, externalists who reject Williams’s explanatory constraint find themselves hypothesizing about how a capacity to be responsive to normative reasons can account for moral motivation.

In sum, the problem of moral motivation has incited adherents of the various positions on practical rationality to show how normative practical reasons can explain moral action. The arguments of chapters 1 and 2 have already put accounts of moral motivation that depend on counterfactual internalism or reasons externalism on the defensive. Nevertheless, I will give all the main views on moral motivation a hearing. Since I’ve argued (in chapter 2) that Williams’s explanatory constraint is a criterion of adequacy for any theory of practical reason, I suggest that the best theory of practical rationality will be the one which supports the best explanation of moral motivation. Furthermore, since these theories explain moral action by citing normative reasons for action, the best account of moral motivation will also furnish the best account of what reason there is to be moral.

In this section, I lay out the alternative views on moral motivation, giving particular attention the way each view conceives of the explanatory relation between practical reasoning and moral conduct. In trying to understand this link, philosophers
look to moral judgment. The starting point is an observation which Connie Rosati has called the basic phenomenon of moral motivation (Rosati, 2006). People are typically motivated to act in accordance with their judgments about right and wrong. When someone judges that it would be morally right to $\varphi$, she is ordinarily motivated to $\varphi$; should she later become convinced that it is wrong to $\varphi$ and right to $\psi$ instead, then she ordinarily ceases to be motivated to $\varphi$ and becomes motivated to $\psi$. Thus, there is a regular connection between moral judgment and moral motivation which has to be accounted for, and philosophers have responded by hypothesizing about various ways in which the motivational force of moral judgments issues from practical reasoning.

The main rival accounts of the relationship between practical reasoning, moral judgment, and moral motivation can be distinguished according to the way in which they respond to two questions:

- **Can cognitive states (in particular, beliefs) independently produce motivations?**
  That is, can a cognitive state produce in an agent a motivation at time $t$, without being accompanied by an affective state which exists prior to $t$?

- **What type of psychological state is a moral judgment?** That is, are moral judgments cognitive (i.e., belief-like) states, or affective (i.e., emotional or desire-like) states?

Both these questions rest on a presumed distinction between cognition and affect. Cognition is usually associated with thinking, believing, reasoning, learning, consciousness, decision-making, and memory. Affect, or emotion, is linked to feeling and
wanting. However, there is no consensus definition of either concept among philosophers and cognitive scientists. This is unfortunate, because in meta-ethics the cognition-affect distinction has long been taken for granted. That said, Jesse Prinz offers a sophisticated account of the distinction (cf. Prinz, 2004, pp. 41-51). According to Prinz,

Cognitive states and processes are those that exploit representations that are under the control of an organism rather than under the control of the environment. A representation is under organismic control if the organism has activated it or maintains it in working memory. A cognitive state is one that contains such a representation, and a cognitive process is one that activates, maintains, or manipulates such a representation (Prinz, 2004, pp. 45 – 46).

Prinz treats “organismic control” as a synonym for “executive” control. And he identifies the “executive” brain structures as those centered in the prefrontal cortex. These structures are thought to play a special role in manipulating and maintaining representations in memory, and in activating representations which are not elicited by the environment (Prinz, 2004, p. 47). The sorts of representations that Prinz takes to be under “organismic control” are concepts, as opposed to percepts. When you see a shape, you have a percept of that shape. But you can also store a copy of that percept in your memory, which you can later on recall. The copy of that percept drawn from memory is a concept of the shape. Thus, a cognitive state, in Prinz’s view, is a mental state which contains concepts (Prinz, 2004, p. 46).

---

7 Even this status quo has begun to unravel. Some contemporary philosophers and emotion researchers argue that cognition and affect are not really distinct psychological operations. Some have argued, for instance, that cognitions can be constituted by emotions, and vice versa (e.g., Lazarus, 1984; Nussbaum, 2001).
Prinz’s definition unifies a lot of widely held intuitions about what mental activities count as “cognitive” (Prinz, 2004, p. 48). For instance, cognition is associated with the conscious processing of information (“access consciousness”). Such processing includes acts of thinking, like doing long division, and deliberation about actions. It therefore it entails organismic control. Also, cognition is thought to be effortful. Choosing a goal for oneself, for instance, seems to require some kind of mental work. The frontal cortex contributes to making strategic choices among competing goals. By contrast, affect occurs unbidden. You don’t have to make an effort to feel sad or angry. Rather, these emotions are induced by events in the surrounding environment.

Prinz reviews evidence for the cognition-affect distinction presented by psychologist R.B. Zajonc. One piece of evidence is the fact that emotion and cognition involve separate neuro-anatomical structures in the brain (Prinz, 2004, pp. 34 – 35). For instance, there is evidence for a subcortical pathway from the retina to the amygdala with no mediation from the neocortex. The neocortex is thought to be the main locus of cognition. This subcortical pathway probably mediates fear responses to dangerous looking objects, like coiled or snake-like things. Prinz suggests that if fear can occur without mediation from the neocortex, then it can occur without cognition. A second line of evidence for the cognition-affect distinction suggests that emotional states can be induced by direct physical means, and not by any prior mental states (Prinz, 2004, p. 35, 37). Emotions can be induced by drugs, hormones, electrical stimulation, and even by facial expression. For example, people who smile more tend to feel happier. In one study,
subjects were asked to evaluate stories with many “ü” sounds, which caused facial
expressions associated with negative emotions (e.g. frowns, grimaces), and nearly
identical stories with many long “ē” and other sounds associated with positive emotions
(e.g., smiles). Most subjects said the “ü” stories were less pleasant than the “ē” stories
(Prinz, 2004, pp. 35 – 36). These findings suggest that emotions do not have to be
prompted by any conscious judgment or complex thought process.

3.2.2 Motivational Humeanism

In the literature on moral motivation, the Humean theory of motivation (or HTM)
seems to be everyone’s foil. Because many important accounts of moral motivation have
been developed in reaction to HTM, it helps to begin my review of the existing
perspectives with Humeanism. Humeans are distinctive for giving an answer of “no” to
the first question in section 3.2.1. They believe it is not possible for a cognitive state to
produce a motivation in an agent, unless it is accompanied by some affective state. To see
precisely what this means, it’s necessary to understand how Humeans explain intentional
action. Intentional action is behavior explained by intentional states (see chapter 2,
section 2.3). The intentional states that figure in Humean explanations of intentional
action are beliefs and desires. It is common practice among Humeans to treat beliefs and
desires as archetypes of cognitive states and affective states, respectively. Because beliefs
are considered cognitive and desires are considered affective, Humeans assume that no
belief is a desire. Humeanism is often called the belief-desire model of action, because it
explains behavior by citing a combination of a desire for some end and a belief that a
certain action is a means to satisfying that desire. The basic Humean position with respect to the first question, then, is that no belief can motivate intentional action without being accompanied by a desire to carry out the action. For example, suppose that while on his morning walk Smith sees storm clouds amass overhead, and he forms the correct belief that it’s about to rain. Luckily, Smith brought his umbrella with him, and through practical reasoning he also acquires the means-ends belief that by opening his umbrella, he will stay dry. According to the Humeans, these beliefs by themselves cannot motivate Smith to do anything. In order for them to play any role in motivating Smith’s action, they must be joined by a desire—for instance, a desire not to get wet. This combination of beliefs and desires is sufficient to produce in Smith a new desire which motivates him to open his umbrella.

A Humean explanation of this sort is subject to a few restrictions. First, the desire not to get wet is necessary for Smith to acquire the desire to open his umbrella; without the former, Smith could not obtain the latter. Second, Humeans claim that the desire not to get wet must exist prior to the time at which Smith acquires a desire to open his umbrella. Also, new desires can be generated in the agent by practical reasoning only if some pre-existing desire is present. Beliefs can contribute to the formation of new desires, but only when joined by some pre-existing desire. To sum up: no belief is a desire, new desires are formed only in the presence of existing desires, and no belief is sufficient to produce a desire.
The second question from section 3.2.1 concerns what sort of psychological state is expressed by a moral judgment. On this question, the Humeans are divided. Some are cognitivists about moral judgment, but others are non-cognitivists (Miller, 2003, p. 3). Cognitivism is the view that a moral judgment expresses a cognitive state—in particular, a belief. What’s important about beliefs is that they are truth-apt, meaning they are capable of being either true or false. Non-cognitivism is the view that a moral judgment expresses a non-cognitive state such as an emotion or desire. Non-cognitive states are supposedly not truth-apt.\(^8\) Desires, for instance, cannot be true or false—they can only be satisfied or unsatisfied. The classic form of non-cognitivism is the emotivism of A.J. Ayer and Charles Stevenson. Emotivism holds that moral judgments are nothing more than expressions of emotion. So, for instance, to express the sentence “Cruelty is wrong” is equivalent to scornfully muttering “Booh on cruelty!” Contemporary versions of non-cognitivism have made advances in explaining how moral judgments can bear inferential relations to each other in a way that purely emotional attitudes apparently cannot. These more sophisticated theories have been called expressivism, with the most prominent expressivists being Simon Blackburn (1984) and Allan Gibbard (1990). Gibbard, for instance, argues that a moral judgment expresses one’s “acceptance” of norms. A norm, in Gibbard’s definition, is “a possible rule or prescription, expressible by an imperative” (Gibbard, 1990, p. 46). Roughly speaking, to “accept” a system of norms is to be

\(^8\) Some philosophers have questioned the widespread assumption that non-cognitive states cannot be true or false. See for example Jesse Prinz, 2007.
emotionally committed to that system. When one violates the norms that one accepts, one is disposed to feel guilty at having violated them, and when one witnesses someone else violating the norms, one is disposed to feel anger toward the offender.

The non-cognitivist Humean’s picture of the relationship between moral judgment and moral motivation is not entirely straightforward. Motivational Humeans conceive of motivations as pairs of beliefs and desires (e.g., my desire to see the Stones concert plus my belief that the Stones concert is in Raleigh tonight motivates me to go to Raleigh). Non-cognitivists conceive of moral judgments as expressions of affective states—particularly emotions. We would like to clarify the relationship between making a moral judgment that expresses an emotional state, on one hand, and having a combination of beliefs and desires sufficient to motivate action in accordance with that judgment, on the other. There are numerous ways to characterize this connection, and I cannot comment on all of them. Instead, I introduce for clarity’s sake a relatively simple and easy-to-understand account of the relationship between emotion and Humean motivation, while admitting that there are other, more nuanced, possibilities. Finn Spicer calls this account the nomological view of the relation between emotion and Humean belief-desire motivation (Spicer, 2004, pp. 56 – 58). It asserts that emotions robustly co-occur with beliefs and desires. This robust co-occurrence is explained by law-like regularities which

---

9 See the illuminating discussions of the relationships holding among emotion, belief, desire, and action in Spicer (2004), Prinz (2004), and Elster (2007).
10 Whatever the true account of this relationship is, a Humean has to maintain her commitment to the belief-desire model of motivation. If motivational Humeanism is true, then insofar as emotions cause or constitute motivation, they must do so in virtue of causing or constituting bundles of beliefs and desires.
govern causal relations among emotions, beliefs, and desires. The regularities guarantee
the truth of conditionals such as “If one is fearful of \( X \), then one believes that \( X \) is a
source of danger and one desires to avoid \( X \).”

On the nomological view, emotions are distinct psychological states from beliefs
and desires. Even so, emotions can motivate action by having effects on one’s beliefs and
desires. For instance, the fear prompted by creaking sounds in my downstairs living room
late at night can make certain beliefs and desires more salient to my mind than others. My
general desire to be safe might spike in intensity. A belief that I could’ve left the deadbolt
on my front door unlatched might occur to me, even though I could’ve sworn that I
locked it. In the grip of these thoughts I might form a desire to check things out. The
point is that an emotion can make “present to mind” a bundle of beliefs and desires that is
effective in motivating me to act in a way that is consistent with the Humean theory of
motivation. Applied to the case of moral motivation, the nomological view suggests that
if a moral judgment expresses an emotion, then a moral judgment can induce a
combination of beliefs and desires that effectively motivates me to act on the judgment.
If, for example, I judge that I have a moral duty to donate blood, and I express my
judgment by the sentence “I morally ought to donate blood,” then on a non-cognitivist
analysis I am expressing an emotional state which can induce a combination of beliefs
and desires that motivates me to donate blood.

Motivational Humeans can also be cognitivists about moral judgment. Cognitivist
Humeans hold that moral judgments express truth-apt beliefs. And as Humeans, they are
committed to the view that a moral judgment can motivate only if it is joined by an independent desire. That is, if agent $S$ judges that she morally ought to do $A$, then on a cognitivist Humean account, $S$ will be motivated to $A$ only if $S$ has some desire to $A$. Furthermore, $S$’s desire to $A$ must be distinct from her moral judgment, and it cannot be produced by her moral judgment alone. For example: if I believe that I morally ought to donate blood, I will be motivated to do so only if I have some desire to donate blood. In addition, my desire to donate blood is not identical to my belief that I ought to donate blood, and my desire is not produced by my belief alone. My belief that I ought to donate blood is, on the cognitivist view, motivationally inert by itself.

One articulate proponent of cognitivist Humeanism is Nick Zangwill (Zangwill, 2003, in SLC, 2007). Zangwill calls his view moral content externalism. It is called “externalist” because Zangwill holds that moral motivation is “external” to moral judgment. This is different from the non-cognitivist Humeans, who argue that moral motivation is constituted by moral judgment and is in that sense “internal” to it. In addition, Zangwill’s position is called “content externalist” because he holds that moral judgments are intentional states that have representational contents. The contents of moral judgments represent “mind-independent moral states of affairs.” Zangwill defines a mind-independent moral state of affairs as “the possession, by a person or their actions or passions, of a real and mind-independent moral property” (Zangwill, 2003, in SLC, p. 244). Moral properties are properties like goodness, justice, virtue, kindness, etc. Typical bearers of moral properties include actions, persons, institutions, policies, practices, and
so on. Now Zangwill is also a moral realist, which means he is committed to the existence of mind-independent moral properties. But it’s also possible for a cognitivist Humean about motivation to be a moral anti-realist who denies the existence of mind-independent moral properties. For instance, one could be an error theorist—a moral anti-realist who maintains a cognitivist interpretation of moral judgment (see Joyce, 2001). In that case, moral judgments would only purport to represent mind-independent moral properties, and in so doing they would always misrepresent. Whether one holds this anti-realist position or Zangwill’s, what unites them is the thesis that moral judgments, if they are to motivate, must be supplemented by independent desires to do as the agent judges she morally ought to do.

In Zangwill’s view, agents are motivated to act on their moral judgments primarily by moral desires. These are desires that contain representations of moral properties (Zangwill, 2003, in SLC, p. 244). An agent will be motivated to act on a moral judgment whenever that judgment is joined by a moral desire that has matching representational content. That is, when S makes a moral judgment that she morally ought to do A in order to achieve some moral state of affairs E, and this judgment is joined by a desire to achieve E, S will be motivated to do A. Thus, according to Zangwill’s moral content externalism, moral motivation “is a matter, first, of believing that some act of ours would have the property of being morally better than some alternative, plus second, there is an independent desire to do actions with that property” (Ibid.).
I think the term “moral desire” is useful for characterizing both cognitivist and non-cognitivist Humeanism. However, I slightly adjust Zangwill’s definition to make it more neutral between cognitivism and non-cognitivism:

*Moral desire*

S has a moral desire to achieve end E just in case S judges that E is a moral state of affairs with certain moral properties, and S desires E.

In short, a moral desire is a desire for whatever the agent judges to be a moral state of affairs. The definition is straightforwardly compatible with Zangwill’s understanding of moral judgments as belief-states and his commitment to mind-independent moral properties. But this definition is also compatible with non-cognitivist Humeanism, because S’s judgment that E is a moral state of affairs with certain moral properties may itself express a moral desire to E. Also, the definition above does not require that E is a *mind-independent* moral state of affairs. It is therefore not prejudiced against at least some forms of non-cognitivism. Expressivists like Simon Blackburn can accept that there are moral states of affairs—i.e., persons, actions, emotions, etc. that can bear moral properties. What they deny is that there are any *mind-independent* moral states of affairs. Blackburn in particular endorses the doctrine of *projectivism*, according to which “evaluative properties are projections of our own sentiments (emotions, reactions, attitudes, commendations)” (Blackburn, 1984, p. 180). Thus, in the above definition of
moral desires, $E$ can be a mind-dependent moral state of affairs constituted by a projection of $S$’s affective states.

Whether one is a cognitivist Humean or a non-cognitivist Humean, moral judgments and moral desires necessarily involve moral concepts. Moral concepts are mental representations of moral properties. Moral judgments and moral desires contain moral concepts. Since, according to Humeanism, moral motivation is explained by moral judgments and moral desires, it follows that any creature capable of moral action must the sort of creature that has moral concepts. This assumption is uncontroversial in the debate about moral motivation. It’s important to highlight this assumption, however, because different theories of moral motivation hold different accounts of the way in which moral concepts contribute to moral motivation. I will argue below that the Humean account of the relation between moral concepts and moral motivation is most plausible.

It’s important to underscore the difference between how cognitivist Humeans explain moral motivation and how the non-cognitivist Humeans explain it. Non-cognitivists hold that moral judgments necessarily entail motivation, since any moral judgment is itself an expression of a non-cognitive motivational state (namely, an emotion or desire). Cognitivist Humeans like Zangwill can acknowledge that there is indeed a tight connection between moral judgment and moral motivation, but they admit that the relationship is contingent in the sense that it can possibly break down. Zangwill, for instance, suggests that people will typically be motivated to act according to their moral judgments, because “it is in some sense natural for human beings to have a
standing general moral desire” (Zangwill, 2003, in SLC, p. 244). This “general moral desire” can be construed as simply a desire to do whatever one judges to be morally good or right. Such a desire, according to Zangwill, will typically accompany a person’s moral judgments, and as a result he or she will typically be motivated to act according to those judgments. However, cognitivist Humeans must allow for the possibility of amoralism—i.e., the possibility that someone can have a moral belief without a corresponding moral desire to act on the belief. Nevertheless, cognitivist Humeans can still explain the basic phenomenon of moral motivation by the hypothesis that moral judgment is reliably correlated with moral motivation. As Zangwill writes, the “bare possibility of someone who has moral beliefs but no moral desires at all is unlikely to be realized, but it is not unthinkable” (Ibid.).

So far we’ve seen how motivational Humeans of different stripes treat moral judgment, and how they explain the basic phenomenon of moral motivation. But we still have to clarify how the Humeans understand the relationship of practical rationality to both moral judgment and motivation. The central idea is this: agents are motivated by instrumental reasons to act in accordance with their moral judgments. Recall from chapter 1 (section 1.1) that an instrumental reason for action is a reason to pursue the optimal means to attaining some desired end. According to the Instrumentalist Dictum, one’s best instrumental reason for action is to do whatever will maximally satisfy one’s desires. For Humeans, then, practical reasoning carried out in conformity with the
Instrumentalist Dictum motivates moral action.\textsuperscript{11} Suppose that, as the non-cognitivist Humeans hold, moral judgments express desire-like states. Then if acting on the desires expressed by one’s moral judgments would maximize the overall satisfaction of one’s desires, the agent’s best instrumental reason is to act on those judgments. Provided that he or she is instrumentally rational, an agent will be motivated to act on this reason. When I judge that I morally ought to donate blood, and if this judgment expresses a desire whose satisfaction would maximize the aggregate satisfaction of my desires, then my best instrumental reason for action is to donate blood. And as long as I’m instrumentally rational, I will be motivated to donate blood. Next, suppose that moral judgments are beliefs, as the cognitivist Humean would have it. In that case, if there is some moral judgment such that there is a moral desire to act as this judgment requires, and the satisfaction of this moral desire would maximize the aggregate satisfaction of the agent’s desires, then the agent’s best instrumental reason will be to act according to the moral judgment. Accordingly, if the agent is instrumentally rational, he or she will be motivated to act on this moral judgment. When I judge that I morally ought to donate blood, and this judgment is joined by a moral desire whose satisfaction would maximize the aggregate satisfaction of all my desires, then my best instrumental reason will be to

\textsuperscript{11} Allan Gibbard, who I have cited as an expressivist, would probably repudiate this claim. Gibbard argues that practical reasoning is not beholden to instrumental means-ends reasoning (cf. Gibbard, 1990, pp. 12 – 18). Instead, Gibbard applies his expressivist analysis to judgments about practical reasons. According to Gibbard, to judge that one has a reason to do $A$ is to express an emotional commitment to a system of norms which permits or requires $A$ (Miller, 2003, p. 96).
donate blood. And as long as I’m instrumentally rational, I will be motivated to donate blood.

From these comments let me summarize the Humean view of the relationship between practical reason, moral judgment, and moral motivation. In chapter 1, I presented the Humean theory of motivation (or HTM) as follows (cf. chapter 1, section 1.4.3):

*The Humean theory of motivation (HTM)*

(H1) No operation of reason is a passion (desire).

(H2) Only the passions (desires) are sufficient to produce motivations.

(H3) The only operation of reasoning that transmits motivational force is reasoning about the optimal means to satisfying a passion (desire).

From (H1) through (H3), I showed that the following thesis can be derived:

(H) Reasoning alone is *not* sufficient to produce a motivation.

This particular formulation of HTM emphasizes the relationship between operations of practical reasoning and an agent’s motivational states (specifically, his or her desires). However, it’s not easy to see how moral judgment can play an explanatory role in motivation. A role for moral judgment in HTM can be set aside, however, if we think of
HTM as a three-part story for the explanation of intentional action. A three-part model, as defined in chapter 2, affirms that a process of practical reasoning can explain action by producing intentional states, and these states in turn motivate intentional action (cf. chapter 2, section 2.4). According to HTM, the only type of practical reasoning capable of generating motivation is procedural. The basic inputs of procedural practical reasoning are the existing motivations of the agent (cf. chapter 1, section 1.3). The existing motivations must be present before any new motivations can be generated by procedural reasoning. In HTM, moral judgment can enter the picture as a causal intermediary between procedural reasoning and motivation. Under HTM, procedural practical reasoning consists in instrumental reasoning about the optimal means to satisfying one’s desired ends. If a moral judgment expresses a desire, then sound practical reasoning will generate means-ends beliefs about how to satisfy that desire. If an agent’s moral judgment expresses a belief, and the agent has a moral desire to act on that belief, then likewise practical reasoning will generate means-ends beliefs about how to satisfy that moral desire. Whenever the satisfaction of a desire to act according to a moral judgment would maximize the aggregate satisfaction of the agent’s desires, the agent’s best reason for action is to act on that judgment. And provided that the agent is instrumentally rational, he or she will form a new desire to act in accordance with his or her best instrumental reasons.

But even means-ends beliefs arrived at through procedural practical reasoning are never by themselves sufficient to generate new desires. Adherents of HTM assume that
for any intentional action $A$ and any agent $S$, the presence of a desire to $A$ is a necessary condition for $S$ to $A$. In addition, Humeans about motivation assume that beliefs and desires are “distinct existences,” which is only to say that no belief is a desire (cf. McNaughton & Rawling, 2004, p. 116; Shafer-Landau, 2003, p. 122). Consequently, proponents of HTM hold that a means-ends belief cannot generate action unless it’s accompanied by a desire to be satisfied. In other words, the combination of agent $S$’s desire to achieve end $E$, together with $S$’s means-ends belief that by doing $A$, he will achieve $E$, is sufficient to generate within $S$ the desire to do $A$. But by itself, a means-ends belief whose content is the statement “$A$ is a means to $E$” is motivationally inert. All these points can be captured in a second formulation of HTM, which I dub HTM*.

\[\text{HTM*}: \text{Second formulation of the Humean theory of motivation (cf. section 1.4)}^{12}\]

\[\text{12 Some philosophers who are widely seen as Humeans about motivation might reject one or more of the five propositions which comprise HTM*. Bernard Williams, in particular, would probably only agree to (H1*). In chapter 1, section 1.4.3, I offer a detailed reconstruction of Williams’s theory of motivation, or (WTM). Williams’s theory still makes procedural reasoning the only type of practical reasoning that can generate motivation, but it conceives (rather vaguely) of procedural reasoning in broader terms that mere means-ends reasoning, and it allows for the existence of motivating states other than desires. The literature is not univocal about what exactly one must believe in order to be considered a “Humean” about motivation. Even more vexing is the fact that philosophers have advanced varying interpretations of motivational Humeanism (for a brief overview, see in particular McNaughton & Rawling, 2004, pp. 116 – 117). But I cannot pursue the matter of what Humeanism “really” amounts to any further. Instead, I present HTM* as is for two reasons: first, it’s most common formulation of Humeanism one encounters in the literature (e.g. in Shafer-Landau, 2003; Rosati, 2006; Korsgaard, 1986; Parfit & Broome, 1997, etc.), and second, it’s the motivational theory that underlies the account of instrumentalism about reasons I presented in chapter 1. I should add that my formulation of HTM* is a bit self-serving, since it’s my aim to show that instrumentalism meets Williams’s explanatory constraint better than rival theories of practical reason.}\]
(H1*) Only procedural practical reasoning is capable of motivating intentional action.

(H2*) Procedural practical reasoning consists of forming justified means-ends beliefs about optimal means for satisfying desired ends.

(H3*) No belief is a desire.

(H4*) No belief is by itself sufficient to motivate intentional action.

(H5*) For any intentional action \( A \) and any agent \( S \), \( S \)'s having a desire to achieve end \( E \) and a means-ends belief that \( A \) is a means to achieving \( E \) is a necessary and sufficient condition for \( S \) to form a desire to \( A \).

3.2.3 Anti-Humeanism

The next several philosophical views on motivation can be grouped under the heading of *anti-Humeanism*. They are united by four commitments concerning the relationship between practical reasoning, moral judgment, and moral motivation. First, all the following anti-Humean theories presuppose moral absolutism. Again, this is the view that moral obligations apply to all individuals, independently of their particular desires or interests. No one can escape a moral obligation simply by not having an interest in doing one’s duty.

Second, the anti-Humean views discussed in this section endorse cognitivism about moral judgment. All of them treat moral judgments as either cognitive psychological states, or as hybrid psychological states with both cognitive and affective
features. In other words, anti-Humeans consider moral judgments to be either beliefs or hybrid states with both belief-like and desire-like properties. Accordingly, anti-Humeans would answer either “yes” or “both” to the second classificatory question posed in section 3.2.1. This has important ramifications for how the anti-Humean treats practical rationality and moral motivation.

The third commitment that unites anti-Humeans about motivation is their rejection of \((H1^*)\) in \(HTM^*\). Anti-Humeans are still interested in providing an account of the basic phenomenon of moral motivation—i.e., the reliable correlation between someone’s judging that she morally ought to \(A\) and her being motivated to \(A\). To this end, anti-Humeans look to moral judgment as the basic source of moral motivation. But this amounts to denying that action can be motivated only by procedural practical reasoning, and therefore to giving a “yes” answer to the first question posed in section 3.2.1. Anti-Humeans allow for moral beliefs, \(qua\) cognitive or semi-cognitive states, to generate a new motivation in an agent at time \(t\) without being accompanied by a desire or other affective state prior to \(t\).

The fourth unifying commitment of the anti-Humeans is that all are moral rationalists—they hold that moral requirements necessarily entail normative reasons for action. As it’s been emphasized before, moral rationalists deny that a person’s reasons for action can be contingent on one’s existing motivations, and this leads moral rationalists to embrace either counterfactual internalism or externalism about reasons. Moral judgments, then, are beliefs about the reasons one has to act. We’ll see below that for some anti-
Humeans (Korsgaard, Nagel, Scanlon, Darwall), it is through practical reasoning that we form judgments about our reasons to act. The type of practical reasoning these anti-Humeans envisage is *substantive reasoning*, by means of which an agent derives conclusions about the reasons for action one has from objective values rather than one’s existing motivations. Other anti-Humeans (e.g., John McDowell, Margaret Little), however, suggest that judgments about one’s reasons for action need not issue from any sort of reasoning at all.

I have characterized Thomas Nagel and Christine Korsgaard both as counterfactual reasons internalists. In other words, they hold that if any agent $S$ has a normative practical reason $R$ to do action $A$, then $S$ has some motivation $M$ to $A$ *provided that* he or she is substantively rational. To be substantively rational is to be motivated by reasons for action entailed by substantive principles of practical rationality (cf. chapter 1, section 1.3). Like actual internalists, counterfactual internalists accept Williams’s explanatory constraint on reasons. Korsgaard is most explicit about this (particularly in her “Skepticism about Practical Reason”), but Nagel endorses Williams’s principle as well (Nagel, 1970, in DGR, p. 333). The question, then, is how counterfactual internalists make sense of the idea that reasons for action derived from substantive principles can motivate human agents. This problem is more acute for the counterfactual internalist than the actual internalist, because substantive practical principles are not grounded in the existing motivations of the agent. Rather they are grounded in what I’ve called objective
values about what is good, right, or worth doing. How, then, can practical reasoning from so-called objective values motivate flesh-and-blood human beings?

Nagel, working in Kant’s shadow, ventures an extended response to this question in *The Possibility of Altruism* (1970). An objective value which Nagel identifies as a source of practical reasons is what he calls a *considerateness for others* (Nagel, 1970, in DGR, p. 335). This is a sensitivity to considerations that favor acting to protect or promote the interests and well-being of other people (Nagel 1970, in DGR, pp. 334 – 35).

Nagel argues that the considerateness for others, and the reasons provided by it, constitute a motivation which can cause rational agents to act altruistically. Altruism, in Nagel’s definition, is “a willingness to act in consideration of the interests of others, without the need of ulterior motives” (Nagel, 1970, in DGR, p. 333). Nagel is careful to distinguish the considerateness for others from affective states—particularly desires—that might dispose someone to act in others’ interests. For Nagel, altruism is not merely an affective penchant for “benevolence, sympathy, love, redirected self-interest,” and so on (Nagel, 1970, in DGR, p.334). Instead, the considerateness for others is a “passionless” motivation because it can be present even when all of the foregoing emotional influences are absent (Ibid.). Therefore, if the considerateness for others is a basic and independent source of motivation, then altruistic action need not be explained ultimately with reference to a desire or emotion. It can instead be explained by principles of substantive practical reasoning. Nagel is plainspoken on this point:

“When action is explained by reasons, it is brought under the control of normative principles. A consideration can operate as a motivating reason only if it has, or is
thought to have, the status of a reason in the system of normative principles by which individuals govern their conduct (Nagel, 1970, in DGR, p. 329)

Nagel rejects a strong distinction between normative and motivating (explanatory) reasons insisted upon by some reasons externalists—in particular the likes of Derek Parfit (cf. chapter 2, section 2.0). Indeed, he thinks that normative reasons for action are explanatory reasons. And in contrast to what actual internalists hold, Nagel contends that an agent’s existing motivations are not necessary conditions for the existence of reasons. For instance, as the instrumentalists see it, an agent can have a reason to \( A \) just in case she has a desire to achieve some end \( E \) and a justified means-ends belief according to which \( A \)-ing is a means to \( E \). Accordingly, the presence of an existing desire is a necessary condition for a reason. But Nagel reverses this picture: normative practical reasons entail desires, rather than deriving from them (Nagel, 1970, in DGR, p. 331). He draws a distinction between what he calls motivated and unmotivated desires. Unmotivated desires are those which “simply assail us,” whereas motivated desires “are arrived at by decision and after deliberation” (Ibid.). As examples of unmotivated desires Nagel cites the appetites, such as hunger, and in certain cases, the emotions. The most salient example of a motivated desire in Nagel’s analysis would be the “desire” to act altruistically. This desire, however, does not exist prior to any normative reason to help others. Nor is this desire a condition for the existence of a reason to act altruistically. It does not even contribute to the explanation of altruistic conduct. Rather, the motivated desire to act altruistically is a logical consequence of one’s normative reason to be considerate of others’ interests. This is the case because Nagel holds a weak definition of
a desire as “whatever may be the motivation for someone’s intentional pursuit of a goal” (Ibid.). For Nagel, a desire is anything that motivates intentional action; it is not any specific type of psychological state—affective, cognitive, or otherwise. Thus, it is just insofar as an agent is motivated to act altruistically we can attribute to him or her a desire for altruism.

What’s most important for Nagel is to understand where the motivated desire for altruism comes from. In his view, it comes from substantive deliberation about one’s reasons for action. By means of this deliberation, one can and ought to recognize that the interests, desires, and needs of others generate practical reasons for oneself. The considerateness for others is a rational constraint on conduct because, as Nagel argues at length, it is a consequence of a “conception of oneself as merely a person among others equally real” (Nagel, 1970, in DGR, p. 329). Nagel contends that we should adopt this conception because practical reasons are by nature impartial or universal (Nagel, 1970, in DGR, pp. 335 – 338). We can see this, Nagel suggests, in the simple fact that you resent others who harm you. Such resent would be baseless, unless the perpetrator was acting contrary to reasons for action he or she had that are supplied by your interests, desires, needs, welfare, and so on. In other words, you take your own welfare to present an objective value that others ought to recognize as a constraint on their conduct, no matter what their particular desires or inclinations might be. But if it’s possible for your welfare to confer reasons for action upon others in this way, then it’s similarly possible for other people’s welfare to confer reasons upon you, regardless of your particular inclinations.
Therefore, Nagel concludes that reasons for action have to be assessed from a general, impartial point of view. If you think that others have reasons to take your welfare into consideration, then you are committed to a general principle which makes no specific reference to you or anyone else. This is a universal principle which tells us that there is someone whose interests and needs all other people have reasons to respect. If this principle did not hold, then you wouldn’t have any basis for resenting others who harm you because there would be no justification for the thought that other people have reasons to take your welfare into consideration. Nagel thus affirms a “requirement of universality on practical principles” which demands that we maintain a considerateness for others and construe ourselves as “one person among others, equally real.”

Accordingly, in Nagel’s picture the motivation to act altruistically derives from the recognition of a reason to be considerate of others’ welfare. This recognition is based on a substantive principle of practical reasoning which requires us to generalize from the self-serving intuition that other people have reasons to respect our own welfare to the universal maxim that there is someone whose welfare all people have reasons to be concerned about. What is, for my purposes, significant about Nagel’s view is that the considerateness for others has motivational force solely because it has normative force, and not because of the relation of this considerateness to an agent’s existing motivations.

In taking this position, Nagel self-consciously follows Kant’s lead:

Kant’s effort to produce a categorical imperative is an attempt to discover requirements on action which apply to a man on no conditions about what he wants, how he feels and so forth. They must nevertheless be requirements whose validity involves the capacity to be motivated in accordance with them…That is,
what makes the requirements valid for us must itself determine the capacity of our motivational structure to yield corresponding action (Nagel, 1970, in DGR, p. 328).

As Nagel (and Kant) would have it, what explains our capacity to be motivated by reasons is practical reasoning itself. Moreover, practical reasoning is substantive, deriving purely from normative considerations rather than from the agent’s own existing motives. This puts into perspective Nagel’s views on moral motivation, as well. For Nagel is both a moral absolutist and a moral rationalist. He thinks that moral requirements apply to all persons, and they necessarily provide reasons for action. Such moral reasons are apprehended in the moral judgments of rational agents who undergo a process of substantive practical reflection. Moral judgments in turn motivate the agent to act according to his or her moral reasons (cf. Nagel, 1970, in DGR, p. 328).

Another version of anti-Humeanism has ties to the reasons externalism of John McDowell. Margaret Olivia Little (1997) describes this view as being based on the idea that “virtue is moral knowledge.” Its proponents see themselves as heirs to Aristotle, and they include some of the main figures responsible for the resurgence of Aristotelian “virtue ethics” in contemporary moral philosophy. The thesis that virtue-is-moral-knowledge is a variant of moral realism (Little, 1997, in SLC, p. 252). It is also an anti-Humean view because it claims (1) that moral judgment is a cognitive (i.e., belief-like) psychological state, and (2) that moral motivation can be generated solely by cognitive/belief-like states.
For instance, McDowell argues that if an agent fails to be motivated by moral requirements, then he or she demonstrates a cognitive failure. On McDowell’s view, the virtuous person is an ideally practically rational agent. In chapter 1 (section 1.5), I explained that McDowell conceives of a practically rational agent as one who “considers matters aright” in his or her practical deliberations (McDowell, 1998, p. 100). Elsewhere, McDowell suggests that to possess virtue is to possess knowledge that enables one to “get things right,” in the sense that he deliberates correctly about his reasons for action (McDowell, 1998, p. 51). For McDowell, reasons for action present themselves as requirements of a situation. He says that the fact that a situation requires a certain sort of behavior is a reason for behaving in that way (Ibid.). He goes on to characterize a sensitivity to one’s practical reasons, or as he calls it, a “reliable sensitivity to a certain sort of requirement that situations impose on behaviour” as a form of knowledge (Ibid., pp. 51 – 52). Moreover, this reliable sensitivity to practical reasons is the kind of knowledge that a virtuous person has. McDowell gives the example of a kind person who knows what it is like to be confronted with a situation which calls for kindness. If having knowledge entails having true beliefs, virtuous persons reliably form correct beliefs about the reasons which favor virtuous action in a range of situations. Non-virtuous persons, in contrast, fail to form correct beliefs about their reasons for action. The failure to form correct beliefs can be classified as a cognitive failure.

Additionally, virtuous people reliably and consistently behave as moral reasons require: a kind person, says McDowell, can be relied upon to behave kindly whenever the
situation demands it. Thus, the virtuous person reliably forms correct and motivationally efficacious beliefs about his reasons for action. To show how this is possible, McDowell draws an analogy between apprehending moral requirements and perception. For McDowell, the virtuous agent is equipped with a special “way of perceiving” which enables him or her to take as salient all the morally relevant features of a situation. Furthermore, the virtuous agent is characterized by a special “way of conceiving” the world. McDowell holds that virtuous agents have a “conception of how to live.” This conception consists of a skill in classifying moral situations as kind, cruel, obligatory, and so on, as well as a disposition to be consistently motivated by moral requirements (Little, 1997, in SLC, pp. 261 – 62).

Thus McDowell is committed to a philosophical psychology which claims that cognition, independently of any affective states, has the power to motivate (Little, 1997, in SLC, p. 252). Virtue is moral knowledge of practical reasons that can motivate the agent to act morally. Indeed, McDowell suggests that action can count as moral or virtuous only if it were motivated by the agent’s correct judgments about his reasons for action (McDowell, 1998, p. 57):

On each of the relevant occasions, the requirement imposed by the situation, and detected by the agent’s sensitivity to such requirements, must exhaust his reason for acting as he does…But the concept of virtue is the concept of a state whose possession accounts for the actions that manifest it. Since that explanatory role is filled by the sensitivity, the sensitivity turns out to be what the virtue is.

In sum, McDowell’s anti-Humean view amounts to this: virtue is moral knowledge, and moral knowledge is having sensitivity to the moral requirements of a situation which are
tantamount to moral reasons for action. Sensitivity to moral reasons can be interpreted, in
my terms, as having correct moral beliefs about one’s reasons for action. These moral
beliefs, in turn, must be able to motivate moral action if there is any such thing as moral
action at all.

McDowell owes us an explanation as to how moral judgments, construed as
cognitive belief-states, can come to motivate persons. Little defends two theories of the
cognitive states that virtue theorists take to be efficacious in motivating moral action. The
first type of state is not even purely cognitive. It is a hybrid psychological state that has
both cognitive (or belief-like) and affective (or desire-like) features. Accordingly, these
states have been dubbed besires (Little, 1997, in SLC, p. 254). Besires are defined with
clarifies this distinction to argue for the Humean thesis that no belief is sufficient to
motivate, but some virtue theorists have appropriated it for their own devices. According
to Smith, beliefs and desires are two kinds of intentional states that can be distinguished
according to their opposed directions of fit (Smith, 1987, 50 – 58). The direction of fit of
an intentional state with the representational content $p$ can be determined by its
counterfactual dependence on evidence that $p$ is not the case (Smith, 1987, p. 56). Beliefs
have a “mind-to-world” direction of fit, because they are the sort of intentional state that
goes out of existence when the subject perceives evidence that the world is not how he
believes it to be. Thus if I believe that it’s 9am on Monday, but my friend informs me that
it’s actually 10, my belief that it’s 9am will be extinguished because it does not correctly
represent the actual time. On the other hand, desires have a “world-to-mind” direction of fit, since the content of a desire is a representation of how the world is to be. Whenever there is a discrepancy between how the world is and how the subject wants the world to be, a desire has the effect of disposing the subject to bring about the satisfaction of his desire (Ibid.). In more schematic terms, if the subject has an intentional state with the content \( p \) which gets extinguished when the subject perceives evidence that \( not-p \), then this state has a mind-to-world fit and is therefore a belief. By contrast, if the subject has an intentional state with content \( p \), and if, upon perceiving evidence that \( not-p \), the subject is disposed to bring it about that \( p \), then this state has a world-to-mind direction of fit. An intentional state that meets this description is therefore a desire.

A besire is an intentional state that has dual directions of fit. Because besires have a mind-to-world direction of fit, they are belief-like. But because they also have a world-to-mind direction of fit, they are desire-like. Besires therefore have both cognitive (belief-like) and affective (desire-like) features. Virtue theorists such as David McNaughton suggest that to be aware of a moral requirement is to be in a “besiring” state. It is a state whose direction of fit is world-to-mind, in the sense that the agent is motivated to change the world to fulfill the requirement. But this state also has a mind-to-world direction of fit, in the sense that the agent is apprehending a truth about the world (McNaughton, 1988, p. 109).

The second kind of cognitive psychological state to which virtue theorists ascribe motivational power is what Little calls a desire-entailing belief. This is a belief whose
possession entails the possession of a desire as a matter of conceptual necessity (Little, 1997, in SLC, p. 254, 259). In other words, a desire-entailing belief cannot exist without the desire it entails. Unlike besires, desire-entailing beliefs are not hybrid states with dual directions of fit. They are ordinary beliefs with mind-to-world direction of fit. And the desires entailed by desire-entailing beliefs are ordinary desires with world-to-mind direction of fit. However, the idea of a desire-entailing belief is contrary to (H4*) and (H5*) of HTM*, because it suggests that a belief unaccompanied by an existing desire state is by itself sufficient to produce a new desire. Although technically a Kantian rather than a virtue theorist, Thomas Nagel is another philosopher who posits desire-entailing moral beliefs. As we saw, Nagel holds that judgments about moral reasons that issue from substantive practical reasoning are *beliefs* that motivate the agent to act as morality requires. As a logical consequence of motivating the agent to act toward a goal, the moral judgment *ipso facto* produces a desire in the agent (Nagel, 1970, in DGR, p. 331 – 332). However, the virtue theorist does not necessarily think that besires or desire-entailing beliefs derive from substantive practical reasoning. John McDowell is a virtue theorist, and recall from chapter 1 (section 1.5) that he denies the assumption that correct judgments about reasons for action issue from successful practical deliberation. Instead, reasons for action exist in virtue of being apprehended by virtuous agents.

I’ve just reviewed in detail two anti-Humean accounts of how one’s judgments about his or her practical reasons—in particular, his or her practical reasons to act morally—can produce moral motivation. Contrary to Humeanism, these accounts treat
moral judgments as either cognitive belief-like psychological states or hybrid states with both belief-like and desire-like features that are sufficient to provide motivation. Further, they are sufficient to produce new motivations independently of any other motivation which exists prior to the new motivation. I shall now turn to two arguments in favor of the Humean account of moral motivation. My remarks are not designed to favor either cognitivist or non-cognitivist Humeanism. Instead, the arguments to come are intended to defend HTM* as it relates to moral motivation. Moral motivation serves as an important test case for deciding between rival theories of practical rationality, because it is in the moral domain the counterfactual internalists and externalists most commonly locate practical reasons which do not depend on any particular agent’s motivations. If Williams’s explanatory constraint (or E-W) is true, as I’ve been assuming throughout this chapter, then the account of practical rationality which supports the most plausible explanation of how and why moral motivation occurs will be all the more persuasive. In the remainder of this chapter, I argue specifically that moral action is best explained by instrumental reasons.

**3.3 For Humeanism: The argument from continuity**

My first defense of Humeanism also undercuts the anti-Humean view. I treat Humeanism and anti-Humeanism as rival theories of motivational psychology. Humeans argue that moral action can flow only from combinations of affective states (in particular, desires) and means-ends beliefs, while anti-Humeans insist that moral action can be explained either by desire-entailing beliefs or “besires.” As such, Humeanism and anti-
Humeanism are subject to the same criteria of adequacy as any hypothesis in the psychology of motivation.

In this section, I suggest that Humeanism meets one theoretical criterion better than anti-Humeanism. This is the standard that an explanatory hypothesis should be consistent with existing well-established theories. The well-entrenched theory I have in mind is evolutionary theory. The theory of evolution is relevant to moral motivation because investigators have discovered that some non-human animals engage in what I call proto-moral behaviors. Human beings are capable of helping and caring for others in morally significant ways. We extend aid to our family members, friends, neighbors, and even strangers, sometimes at great sacrifice to ourselves. And when we ask ourselves why we help others the way we do, we often resort to moral language—we speak of the duties of a good mother or son, the virtues of a loyal friend, and so on. Thus, some helping and caring behaviors, at least when carried out by human beings, are tokens of moral action. But some non-human animals engage in the same sorts of helping and caring to which we attribute moral import. Such behavior can be called proto-moral, in the sense that it may furnish insight into the evolutionary precursors of an important class of morally significant actions found in humans. Indeed, the fact that the great apes and other social animals are capable of proto-moral behavior suggests that the psychological capacities associated with moral motivation trace their origins back to the ancestors that homo sapiens shares with other creatures.
Evolutionary theory counsels us to postulate psychological mechanisms which can be explained as the outcome of a slow, incremental process of “descent with modification.” Therefore, when we hypothesize about the mechanisms underlying moral motivation, we should be able to tell a story about how the proto-moral motivations found in animals could have been shaped by natural selection into the full-fledged moral motivations of humans. I argue that Humeanism is much better-suited to this task than anti-Humeanism, because the motivations which Humeans posit to account for moral action are the same as, or very similar to, the motivations that researchers have invoked to explain proto-moral helping and caring in certain animals.

For several decades now, researchers in various fields have pursued the task of explaining the evolutionary origins of morality. The guiding assumption of this research program is that the moral sense—that is, the capacity for moral thought and action—is an evolutionary adaptation. The basic questions of the program include: What adaptive problems was a moral sense selected to solve? What fitness-enhancing functions did morality serve in environments inhabited by our human forebears? To answer these questions, some investigators have inquired into the moral or proto-moral capacities of animals. Primatologist Frans de Waal has observed compelling parallels between helping/caring behaviors in human beings and non-human primates. He argues that some primates—particularly the great apes—are motivated by the same psychological mechanisms which underlie the human propensities for empathy and altruism (De Waal, 13 For reviews, see Krebs (2005), Krebs & Janicki (2004), and Joyce (2006, ch. 1).
2006, 2008). De Waal characterizes these propensities as necessary conditions for morality. Furthermore, he holds that empathy and altruism are fundamentally affective motivations based on an “emotional interest in others” (De Waal, 2006, pp. 20 – 21). Since motivational Humeanism posits that moral action is motivated by affective or desire-like states, it finds support in DeWaal’s theory of proto-moral behavior in non-humans. The most important lesson to take from de Waal’s work is the way in which he defends his affective account of empathy and altruism. In particular, when discussing the explanatory role of these motivations in human behavior, de Waal asserts that “[i]t is reasonable to assume that the altruistic and caring responses of other animals, especially mammals, rest on similar mechanisms” (De Waal, 2006, p. 28). It will be my concern to show that this assumption is well-founded.

To understand de Waal’s theory of empathy, we have to draw several distinctions. First is the distinction between behavior and motivation. Behavior refers simply to what a creature can be observed doing, while motivation is a theoretical concept used to explain behavior. Next, we must distinguish between different types of altruistic behavior. Altruistic behavior is, broadly speaking, behavior that somehow helps or benefits another individual (cf. Joyce, 2006, ch. 1). An important topic in contemporary evolutionary biology is the origin of biologically altruistic behavior (see especially Sober & Wilson, 1998). Biologically altruistic behavior is helping behavior that has the effect of increasing the recipient’s fitness (i.e. reproductive success) at a cost to the performer’s fitness. An example of biologically altruistic behavior is the honeybee that stings an intruder,
sacrificing her own life to protect the hive and its queen (De Waal, 2008, p. 280; Joyce, 2006, p. 20). The second type of altruistic behavior is what de Waal calls directed altruism. This is behavior which has the effect of helping an individual in need, pain, or distress (De Waal, 2008, p. 281). Directed altruistic behavior may not be biologically altruistic, and biologically altruistic behavior may not exhibit directed altruism. When a father helps his son pay tuition for medical school, despite the fact that the boy could be spending more of his time getting women pregnant than studying, the father helps his son’s needs but does not enhance his own fitness. And, to use Joyce’s example, when a son kills his elderly mother to free up food and other resources that can be used to sustain more of his own children, he may be advancing his victim’s fitness, but certainly not her needs (Joyce, 2006, p. 15)!

De Waal offers an analysis of the motivational mechanisms which explain directed altruistic behavior. One way in which such behavior could be explained is by the intentional states of an animal. De Waal calls the suite of mechanisms which explains directed altruistic behavior intentional altruism (De Waal, 2008, p. 281).\textsuperscript{14} Intentional altruism is helping behavior motivated by the agent’s intentional states. Among the relevant intentional states that motivate intentional altruism are the agent’s predictions about the effects of extending help. Non-intentional altruism, by contrast, is not motivated by an intentional representation of the future effects of one’s helping. Since

\textsuperscript{14} Intentional altruism has also been dubbed psychological altruism by Sober and Wilson (1998) and Kitcher (2006). Joyce (2006) confines the term “altruism” to mean intentional/psychological altruism.
insects are not generally believed to have intentional states, the honeybee’s suicidal sting would also be a prime example of non-intentional altruism (Okrent, 2007, ch. 5; Carruthers, 2004). In this case, some non-intentional mechanism motivates the bee to carry out a biologically altruistic behavior.

Two forms of intentional altruism can be distinguished according to a distinction between *self-regarding desires* and *other-regarding desires*. Agent $S$ has a self-regarding desire to achieve some end $E$ if $E$ is a state of $S$. By contrast, $S$ has an other-regarding desire for $E$ if $E$ is a state of some other agent (or group of agents), $S^*$.\(^{15}\) For instance, suppose that I donate blood purely from a desire to win the respect of my peers through conspicuous acts of benevolence. If you were to ask me, I would truly admit that if donating blood did not do wonders for my reputation, I wouldn’t have any inclination to do it. In this case, I have a self-regarding desire to donate blood, where the desired end is an improved reputation. An improved reputation would be a state in which I find myself; it is not a state of anyone else. On the other hand, suppose instead that I donate blood because of a desire to promote the health of people I’ve never met. Here my desired end is to change the states that other people are in, and not my own state, by improving the health of those in need of blood transfusions. To the extent that my end is exclusively to improve the health of others, and not to advance my reputation or change my own state in any other way, I have an other-regarding desire to donate blood. The distinction between

\(^{15}\) These definitions are inspired by Joel Feinberg’s distinction between “self-regarding” and “other-regarding” motives in his “Psychological Egoism” (cf. Feinberg, in SL, 2007, p. 190).
self- and other-regarding desires allows for the possibility that someone could be motivated to do something by both self- and other-regarding desires. For instance, it may be that only the combination of a self-regarding desire for improved reputation and an other-regarding desire to promote the health of others is sufficient to motivate me to donate blood. At the same time, each desire might be individually necessary to motivate me. Intentional altruism which operates through self-regarding desires may be called self-regarding altruism, whereas intentional altruism which operates through other-regarding desires can be called other-regarding altruism.

De Waal’s main contention is that empathy is the proximate mechanism that motivates directed altruistic behavior (De Waal, 2008, p. 282). Empathy is defined by de Waal as “the capacity to (a) be affected by and share the emotional state of another, (b) assess the reasons for the other’s state, and (c) identify with the other, adopting his or her perspective” (De Waal, 2008, p. 281). For De Waal, each of these three capacities is individually sufficient to be called empathy. Many animals are not capable of meeting (a), (b), and (c) all at once, but even if an animal meets just one criterion, the term “empathy” still applies (De Waal, 2008, p. 281). De Waal holds that empathy is the primary causal engine through which intentional altruism operates. However, empathy-based altruism really cuts across the distinction between intentional and non-intentional altruism. For instance, an entire flock of birds might take to the air because just one of them was startled. The rest of the flock responds empathetically to the startled individual, since the startled one’s fear is matched by a similar state in the other birds (cf. De Waal,
2006, p. 283). But the flight reaction might be triggered in each individual by a biologically adaptive reflex, rather than an intentional state. Additionally, empathy-based altruism cuts across the distinction between self-regarding and other-regarding altruism. Some of the earliest evidence for empathy-based altruism in monkeys came from a pioneering series of experiments in which rhesus monkeys refused to pull a chain that would deliver food to themselves and an electric shock to a companion. Two of the monkeys refused to pull the lever for so long that they were literally starving themselves (De Waal, 2006, p. 29). It’s unclear whether these results are to be explained by self-regarding or other-regarding altruism, but whichever is the case, empathy may very well play a role. One possibility is that the subject experiences personal distress to match the distress of the conspecific, and becomes averse to pulling the level as a means to sparing itself further internal discomfort. In that case the monkey’s behavior would be motivated by self-regarding altruism, since it experiences distress induced by another monkey but acts on a self-regarding desire to alleviate its own distress. Another possibility is that the subject experiences distress in reaction to the distress of the other, but refuses to pull the lever as a means of alleviating the distress of the other. In this alternative scenario, the monkey would be motivated by other-regarding altruism. Either way, the subject shares in the emotional state of the other, which satisfies one criterion for empathetic concern (namely, criterion (a)). According to de Waal, more recent ethological evidence has allowed us to parse behaviors explained by self-regarding altruism from those best explained by other-regarding altruism.

230
De Waal distinguishes three levels of empathy. The first level is called *emotional contagion*. It consists of the matching of emotional states between one creature (the “subject”) and another (the “object”) (De Waal, 2008, p. 282). A creature capable of emotional contagion meets criterion \((a)\) in de Waal’s definition of empathy. For example, a single crying newborn may prompt a whole room full of newborns to start crying. The distress of one infant “spreads” to the others. Likewise, rats and pigeons display distress in response to perceived distress signals in conspecifics. Mice exhibit intensified pain responses when they see other mice in pain. Also, rhesus macaque monkeys tend to terminate projected pictures of conspecifics in a fearful pose (De Waal, 2008, p. 283, 288).

The second level of empathy is *sympathetic concern*, which involves emotional concern about a distressed or needy other. The sympathizing subject’s concern motivates attempts to ameliorate the distress of the other (Ibid.). Sympathetic concern is to be distinguished from *personal distress*. The latter involves distress induced by the distress of the other, which in turn motivates the affected party to alleviate its *own* distress (Ibid.). Personal distress is the affective state through which self-regarding desires motivate self-regarding altruism. For example, the screams of a punished infant rhesus monkey will cause other infants to embrace or mount the distressed party. The distress of one infant spreads by emotional contagion to its peers, who seek to reduce their own negative affect (De Waal, 2006, p. 27). Sympathetic concern, on the other hand, is the affective disposition through which other-regarding desires motivate other-regarding altruism. A
compelling manifestation of sympathetic concern is consolation behavior, which de Waal defines as “reassurance provided by an uninvolved bystander to one of the combatants in a preceding aggressive incident” (De Waal, 2006, p. 33). For example, de Waal observed a juvenile chimpanzee approach and puts its arms around a screaming adult male who had just been defeated in a fight. Quantitative studies have found that in chimpanzees, third parties direct comforting behavior more at recipients of aggression than at aggressors, and more at recipients of intense aggression than mild aggression (De Waal, 2006, p. 35). There are important differences between personal distress and the mechanism which motivates consolation behavior. Consolation behavior is more clearly modulated to the state of the other—hence why it is more frequently directed at recipients of intense aggression then recipients of mild aggression or at aggressors.

De Waal’s third level of empathy is exhibited by only a few animal species other than humans—particularly apes, elephants, and cetaceans. This is empathetic perspective-taking (henceforth “EPT”). EPT consists in “the capacity to take another’s perspective—e.g., understanding another’s specific situation and needs separate from one’s own—combined with vicarious emotional arousal” (De Waal, 2008, p. 285). Like the other forms of empathy, EPT involves emotional engagement with the other, but it also requires the cognitive ability to attribute mental states to the other. The best evidence for EPT comes from the phenomenon of targeting helping, which is helping behavior that is sensitive to the specific needs and situation of the other (De Waal, 2006, p. 285). According to de Waal, there are hundreds of qualitative examples of targeted helping in
apes, dolphins, and elephants. Researchers have frequently observed that when juvenile orangutans get stuck in a tree, they are rescued by their mothers. Mother orangutans have been seen draping their bodies between one branch and another, thereby creating a bridge to safety for their whimpering offspring. This behavior requires an understanding of the causes of the juvenile’s emotional distress, as well as an ability to identify a solution to the problem. To take another dramatic anecdote, a captive chimpanzee by the name of Krom failed to remove a tire filled with water that was hung with other tires in a row on a horizontal log. Eventually, Krom gave up and walked away. But then Jakie, another chimp that Krom had cared for while he was a juvenile, approached the log and pushed the tires off it one by one. When he reached the last tire, he removed it without spilling any of the water, and carried it to Krom. As de Waal describes the story’s end, “Krom accepted his present without any special acknowledgment, and was already scooping up water with her hand when Jakie left” (De Waal, 2006, p. 32; 2008, p. 290). Here it’s difficult to see what could be motivating Jakie, except for an other-regarding desire to satisfy Krom’s perceived wishes. It seems that no self-regarding desire of Jakie’s stood to be satisfied by delivering the tire, since Krom was a low-ranking adult female who could not do much to help Jakie in return, and since there were no onlookers present from which he might accrue higher social status through conspicuous acts of helpfulness (Kitcher, 2006, p. 130). The cognitive element in Jakie’s action is also salient, since Jakie would have needed to attribute to Krom a desire for the tire. It seems that Jakie could
only act the way he did if he possessed the ability to take Krom’s perspective and understand what she was going for.

De Waal assumes that the same motivational mechanisms which underlie empathy and intentional altruism in apes and other social animals are also present in humans. What makes this assumption reasonable, I suggest, is a constraint on hypothesis-making in psychology imposed by evolutionary theory. This I call the *continuity constraint*. Organic structures and mechanisms are products of a slow, incremental process of descent with modification. Newer, more complex structures are piecemeal modifications of older, often (but not always) simpler ones. For example: the frontal fins of fish evolved into the front limbs of land animals, which became hoofs, paws, wings, and hands (De Waal, 2006, p. 21). Given that human beings are evolved creatures, the continuity constraint advises us to prefer an account of human psychology which can explain how complex human traits evolved from simpler antecedents resembling the traits of non-human animals. Indeed, Darwin himself argued for the psychological continuity of humans and non-humans (cf. Sober, 1998, p. 228). In the second edition of *The Descent of Man* (1874), he wrote (Thompson, 1995, p. 71):

…the difference in mind between man and the higher animals, great as it is, certainly is one of degree and not of kind…the senses and intuitions, the various emotions and faculties, such as love, memory, attention, curiosity, imitation,

---

16 Evolutionary change does not *only* proceed from the simple to the complex. As Elliot Sober notes, “the history of life is peppered with cases of evolutionary simplification. For example, the evolution of parasites typically involves a transition from complex to simple, as the parasite loses features of its free-living ancestor” (Sober, 1998, p. 225).
reason, &c., of which man boasts, may be found in an incipient, or even sometimes in a well-developed condition, in the lower animals.

Similarly, Darwin believed that the rudiments of human moral psychology could be found in other social animals (Thompson, 1995, pp. 41 – 42):

…any animal whatever, endowed with wellmarked social instincts, the parental and filial affections being here included, would inevitably acquire a moral sense or conscience, as soon as its intellectual powers had become as well, or nearly as well developed, as in man.

Daniel C. Dennett also defends this commitment to phylogenic continuity. Dennett, however, is careful to note that the continuity constraint applies even when a species of organism exhibits seemingly unique and extraordinary traits, as so often seems to be the case with humans. Even when the emergence of a new organism or biological structure at some moment in evolutionary history seems radically different from what existed before, Dennett insists that theorists should search for “a subprocess or special feature of a design process that can be demonstrated to permit the local speeding up of the basic, slow process of natural selection, and that can be demonstrated to be itself the predictable (or retrospectively explicable) product of the basic process” (Dennett, 1995, p. 76). In sum, the continuity constraint requires that whenever a psychological mechanism is postulated, its origins should be explicable in a manner consistent with evolutionary theory. But this theory favors explaining the origins of a trait as a product of descent with modification from pre-existing traits. Since evolutionary theory suggests that humans and other animals—especially the apes—share common ancestors, the older traits on which natural
De Waal proposes a “Russian Doll Model” of empathy which accords with the continuity constraint. Emotional contagion and its neural-physiological realization, which de Waal calls a Perception Action Mechanism (PAM), constitute the “core” of the empathy (De Waal, 2008, pp. 287 – 88). The next “layer” of de Waal’s model is sympathetic concern, and the outermost “layer” is empathetic perspective-taking. De Waal holds that the relatively complex mechanisms and cognitive abilities associated with empathetic perspective-taking evolved through processes of reciprocity and kin selection from simpler mechanisms related to emotional contagion. This explains why monkeys seem to possess emotional contagion, but their capacity for targeted helping is not nearly as robust as has been observed in great apes. For instance, first-time mother macaques sometimes accidentally drown their own infants while bathing themselves (De Waal, 2006, p. 40). According to de Waal, the capacity for empathetic perspective-taking would have been a relatively new phenomenon that emerged with the common ancestor of humans and apes, whereas the capacity for emotional contagion was inherent in our common ancestor with monkeys and probably in much older organisms. At the same time, De Waal argues that all three levels of empathy are operative in creatures capable of intentional altruism. He hypothesizes that the more primitive levels of empathy provide enabling conditions for the more complex ones (De Waal, 2008, p. 287). In order for EPT to work, there has to be a way for an organism to detect and be motivated to alleviate the selection operated to produce the psychological capacities of modern humans can probably be found in nonhuman creatures.
distress of a conspecific. Only then can the organism perceive the source of the distress and attempt to eliminate it. Emotional contagion and sympathetic concern could be effective in making the organism share in the emotional state of the distressed subject. The work left for EPT, then, would be to identify the cause of the subject’s distress and to determine how to alleviate it.

The mechanisms that De Waal theorizes to explain altruistic behavior in animals, and ultimately in humans as well, are consistent with the Humean theory of motivation. A chimp with the capacity for empathetic perspective-taking shares in the emotional state of the one in need or distress, assesses the causes of the need or distress, and then takes whatever means are necessary to meet the need or alleviate the distress of the other. Such a chimp can be said to be motivated by instrumental reasons for action (provided that we accept the non-linguistic animals are capable of acting for reasons—see chapter 2, section 2.3). His instrumental reason consists in his belief that a conspecific is in need, and his desire to meet that need. This belief-desire pair constitutes both an instrumental reason to help the other, and a motivation to act accordingly. In sum, the propensity to share in emotional experience, the ability to attribute mental states to others, and the causal reasoning necessary to understand how to meet others’ needs, seem altogether sufficient for non-human animals to engage in proto-moral helping and caring behavior.

By contrast, de Waal’s Russian Doll Model is not consistent with the motivational mechanisms postulated by anti-Humeanism. For anti-Humean mechanisms are far too cognitively demanding for non-humans to be motivated by them. Consider the virtue
theorist’s account of moral motivation. On this view, a virtuous agent is one who has
developed a “conception of how to live.” We saw in section 3.2 that agents with a
conception of how to live are supposed to be skilled at making moral classifications. This
involves an ability to correctly apply moral concepts such as kind, cruel, and obligatory
to situations that have the appropriate moral properties. Virtuous agents are then
motivated to act in accordance with what he or she perceives as the moral requirements
of a situation. However, there is no persuasive evidence to suggest that non-humans have
even rudimentary moral concepts.\textsuperscript{17} And even if they did, it seems beyond dispute that
non-human animals cannot form a complete “conception of how to live” which enables
them to be consistently and reliably motivated by the moral requirements posed by a
situation. It’s quite implausible that besires or desire-entailing beliefs, characterized in
any way resembling the virtue theorist’s account of them, could have a place in the
psychologies of even the most intelligent non-human animals.

Next, consider Nagel’s Kantian picture of moral motivation. It seems that only
humans have the deliberative powers necessary to understand themselves as “one person
among others, equally real,” and to see how this entails that the interests of others are
reasons for action. Are we to suggest that Jakie comes to Krom’s aid because he
recognizes that, were the situation reversed, he would think that Krom had a reason to

\textsuperscript{17} Some of de Waal’s experimental findings suggest that Capuchin monkeys may have a primitive concept
of fairness. For instance, when one monkey is given a cucumber and his cage mate is given a more
desirable grape, the first monkey often reacts negatively, e.g., by throwing it out of the cage (cf. De Waal,
2006, pp. 44 – 49). However, Kitcher and others point out that these negative reactions could have been
elicted by the mere observed possibility of a superior reward, and not by the perception of unfairness
(Kitcher, 2006, p. 131).
assist him? De Waal’s research suggests that chimps likely are capable of taking the perspective of others. This is undoubtedly a cognitive achievement, but the question at issue is what motivates Jakie to help Krom. If we follow Nagel, it would have to be a desire or desire-entailing belief derived from a passionless considerateness for others—a purely cognitive recognition, from an impartial point of view, that Krom’s needs are objective values that have no less normative force than Jakie’s own needs would in the same situation. Not even an ape has the capacity to engage in the kind of impartial and normatively infused substantive reasoning which Nagel believes to be the source of moral motivation.

The continuity constraint teaches us to minimize discontinuities between human and non-human psychologies. But where such discontinuities are necessary, it teaches us to show that they are products of some evolutionary process. Anti-Humeanism postulates motivational mechanisms that are far more discontinuous from animal motivations than Humean mechanisms. As I will argue in the next section, it is far from obvious that such great departures from animal motivations are necessary to explain moral action in humans. Of course, any theory of moral motivation will have to posit some mechanisms which are unique to human beings. In particular, we should expect a mechanism through which moral concepts can play a role in generating moral action. Every theory of moral motivation surveyed above assumes that moral behavior is controlled by moral judgment, or by the conjunction of moral judgment and moral desire. Moral judgments and desires by definition involve moral concepts. So whatever the true theory of moral motivation
turns out to be, it will have to include an evolutionary story about how human beings developed moral concepts, and how these concepts could be motivationally effective. But this will be a much taller order for anti-Humeanism than Humeanism. For the Humeans overlay moral concepts upon the same mechanisms believed to move animals. Intentional behavior in humans, as in animals, is driven by the beliefs and desires of the agent. The problem for the Humeans is to show how moral concepts come to be contained in the human animal’s beliefs and desires. By contrast, the anti-Humean’s explanatory burden is far more daunting. In addition to explaining the evolutionary origins of moral concepts, they have account for the emergence of an entirely new behavioral engine which is only to be found in the human psyche. This engine consists of some kind of cognitive or hybrid cognitive-affective state—i.e., a desire-entailing belief or besire—whose presence is independently sufficient to cause action. I don’t intend to spout any prophesies claiming that this explanatory burden can never be discharged. But I will suggest that it is unlikely to be done. The anti-Humean would have to tell us what evolutionary pressures would favor besires or desire-entailing beliefs over the belief-desire pairs that motivated our more ape-like ancestors. In particular, he would have to explain why those same evolutionary pressures could not just as well favor beliefs and desires containing moral concepts. If morality is an adaptation, why can’t adaptive moral behavior in humans flow from a relatively slight modification of the same mechanism comprising beliefs, desires, and means-ends reasoning that generates proto-moral behaviors in animals? Why assume that moral behavior has to issue from a radically new and different mechanism? In the
absence of a principled answer to these questions, motivational Humeanism clearly has the upper hand. For the preceding comments indicate that Humeanism stands a far better chance of being included in an evolutionarily continuous etiology of moral motivation.

3.4 For Humeanism: The argument from Morgan’s Canon

My second argument in favor of motivational Humeanism begins with a question: Are there any actions that can be explained by anti-Humeans, but not by Humeans? We saw in the last section that animals are capable of many behaviors which we construe as having moral significance, at least when they are carried out by humans. This led us to wonder whether the mechanisms that underlie moral motivation are specific to humans, or rather if the same or similar mechanisms bring about motivation in other animals. De Waal argues that evolutionary theory supports a presumption in favor of the continuity of human and animal psychology. Others argue that even when discontinuity is observed, we should count on evolutionary theory to explain its occurrence. However, human beings seem to be unique among the animals in having the ability to make moral judgments. Philip Kitcher aptly notes that Darwin’s phrase, “descent with modification,” really captures two countervailing tendencies in the evolutionary process: both continuity and discontinuity with the past (Kitcher, 2006, p. 124). An objector may charge that the argument from continuity elides too much of what makes humanity special—it overemphasizes the extent to which homo sapiens is descended from apes while it underplays the extent to which we are extraordinary modifications of our ape-like ancestors.
But I think another criterion for choosing hypotheses supports motivational Humeanism. This principle is known as Morgan’s Canon, named for the comparative psychologist C. Lloyd Morgan. In his *Introduction to Comparative Psychology* (1894), Morgan states his rule as follows: “In no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of the exercises of one which stands lower in the psychological scale” (quoted in Sober, 1998, p. 224). Morgan’s Canon applies to hypotheses that claim a causal relationship between the specific internal mechanisms of a creature and its behavior. It’s quite appropriate, therefore, to apply Morgan’s Canon to theories of moral motivation. Although Morgan himself thought his principle to be justified by Darwinian considerations, Elliot Sober (1998) argues that it can be substantiated by deductive logic. To show this, Sober offers a novel interpretation of Morgan’s Canon, summarized as follows (Sober, 1998, pp. 236 – 237):

*Morgan’s Canon (as interpreted by Elliot Sober, 1998)*

Let one internal mechanism, $H$, be *higher* than another, $L$, if and only if the behavioral capacities entailed by $H$ properly include the behavioral capacities entailed by $L$. Suppose that $L$ entails the behaviors in set $B_1$ and $H$ entails the behaviors in both sets $B_1$ and $B_2$ (where $B_1$ and $B_2$ are not equivalent sets). Then if an organism performs the behaviors in $B_1$, but not the behaviors in $B_2$, the organism lacks $H$. 

242
On Sober’s interpretation of Morgan’s Canon, “higher” mechanisms are so named because they allow an organism to do more than “lower” ones. Because $L$ “entails” or explains $B1$, and $H$ “entails” both $B1$ and $B2$, the behaviors entailed by $H$ “properly include” the behaviors entailed by $L$. Therefore, $H$ is a higher mechanism than $L$. This means that an organism with $H$ can perform the behaviors in both sets $B1$ and $B2$. However, if an organism has $L$ but lacks $H$, then it can only do the behaviors in $B1$. So construed, Morgan’s Canon states that attributions of psychological mechanisms are constrained by the following deductive inference:

(1) Mechanism $H$ entails the behaviors in $B1$ and $B2$.

(2) Organism $O$ does not display the behaviors in $B2$.

(3) Therefore (by modus tollens), $O$ does not have $H$.

Thus, Morgan’s Canon advises us to attribute a psychological mechanism to an organism only if it’s the sole mechanism which can cause some behavior. But this is not to say that all behaviors must be explained by one and the same mechanism. When a person rapidly opens and shuts her eyes, her behavior might be caused by more than one mechanism; it could just be due to her normal blinking reflex, or it could be explained by her intention to flirt with someone. But Morgan’s Canon would not require us to dispense with intentionality, as long as it can explain other behaviors which the blinking reflex and
other psychological mechanisms cannot. The same considerations apply when different psychological mechanisms are invoked to explain the same behaviors exhibited by different species. To use Sober’s example, stickleback fish and chimpanzees both build nests, but Morgan’s Canon would not then require that the same psychological process causes nest-building in both organisms (Sober, 1998, p. 230). Instead, the Canon can allow the psychological process which motivates nest-building in chimps to be very different from the one which motivates it in sticklebacks, but only if this process *alone* could explain additional behaviors exhibited by chimps but not stickelbacks.

Before Morgan’s Canon can be applied, it must be assumed that the “higher” mechanism $H$ and the “lower” mechanism $L$ entail *different* behaviors. Thus, $B_1$ and $B_2$ must not be identical sets of behaviors. This assumption reflects the idea that any causal mechanism hypothesized to explain a phenomenon must figure *ineliminably* in that explanation. If $B_1$ and $B_2$ designated the same set of observed behaviors, there would be no basis for postulating different psychological mechanisms to explain them. On the other hand if there is some behavior which no known mechanism can explain, then this suggests in favor of hypothesizing an additional mechanism.

It is fair to say that Morgan’s Canon is a principle of theoretical parsimony. According to Sober, however, the Canon must not be confused with a more familiar parsimony criterion: Occam’s Razor (Sober, 1998, pp. 229 – 231). Occam’s Razor requires us to accept the simplest hypothesis compatible with the evidence. But simplicity is not what leads us away from attributing $H$ to an organism that does only $B_1$. Rather, it
is the non-occurrence of $B_2$. Since $H$ entails both $B_1$ and $B_2$, the non-occurrence of $B_2$
would imply by *modus tollens* that the organism does not have $H$ (Sober, 1998, p. 240).\(^{18}\)

Sober offers an illuminating example of how Morgan’s Canon is applied to theory
choice in comparative psychology. Consider a species of bird call the piping plover.
When a piping plover sees a predator approach its nest, it responds with a “broken wing
display,” dragging its intact wing along the sand as if it were injured (Sober, 1998, p. 224). On other occasions, plovers have been seen to engage in “false nesting.” When a
predator approaches, the plover moves away from the nest, sits down, and starts peeping
loudly. In effect, the bird makes it look like it’s sitting on eggs, when in fact it’s sitting on
sand (Sober, 1998, p. 238). What explains these curious behaviors? Consider two
possibilities: Let $H_1$ be the hypothesis which says that the birds want to protect their
young, and they believe these actions will have that effect (without knowing exactly
*how*). Let $H_2$ be the hypothesis that the plovers want to protect their young, believe that
deceiving the predator will accomplish this, *and* they believe that behaviors like the
broken wing display and false nesting will successfully deceive the predator. $H_2$
attributes what philosophers sometimes call *second-order intentionality* to the plovers:
that is, they have the capacity to form representations about the mental states of other

\(^{18}\) Sober also suggests that if we observe an organism doing the behaviors in $B_1$, but not in $B_2$, then
we can make a deductive inference to the conclusion that the organism lacks $H$ but *does* have $L$. But that
would only be the case if we assume that $B_1$ entails $L$. This assumption is unrealistic. The presence of a
behavior does not guarantee that it was caused by any particular mechanism. Instead, the inference to $L$
from the presence of $B_1$ is likely a (non-deductive) inference to the best explanation. Of course, such an
inference may nevertheless be quite reasonable.
organisms. For example, if the plover has second-order intentionality, it believes that the false nesting display will make an approaching predator believe that the plover is sitting on its eggs, when really it’s not. (In the previous section, we saw that De Waal thinks empathetic perspective-taking involves second-order intentionality.)

Morgan’s Canon instructs us to reject $H_2$ in favor of $H_1$. The reason is that the plover is not displaying any behaviors that are explained only by second-order intentionality. Both the broken-wing display and the false nesting are compatible with $H_1$, the hypothesis of first-order intentionality. According to this hypothesis, the plover’s actions can be explained by its intentional states about its own situation—viz., its desire to protect its young plus its belief that the relevant actions will meet this desire—and not intentional states about the intentional states of others. On the other hand, the plover does not exhibit any behaviors which are specific to second-order intentionality, such as targeted helping.

Sober is careful to point out that a large body of observations is necessary to establish that the plover is not displaying the behaviors entailed by a mechanism. For instance, in evaluating whether plovers are capable of second-order intentionality, they should be observed over a long period of time and under varied circumstances (Sober, 1998, pp. 237 – 38). If no plover can be observed behaving in a way which is explained by second-order intentionality only, then none of the plover’s actions should be explained by second-order intentionality.
Morgan’s Canon can be recruited in an argument against anti-Humeanism. The argument begins with an observation that few would think to dispute (see, section 3.5). It would seem that the Humean belief-desire model can successfully explain at least some human behaviors. When explaining the actions of other people, we often appeal to the beliefs and desires that motivated them. Philosophers like to call this explanatory strategy *folk psychology*.\(^{19}\) Folk psychology is commonly deployed in explanations of *non-moral* action.\(^{20}\) By “non-moral action,” I just mean action which is not morally significant. Take eating for example. Given that I’m generally well-fed, I am under no *moral* obligation not to skip a meal. And yet it is not morally wrong for me to eat when hungry, either. Folk psychology provides a very commonsensical explanation for eating. Hunger induces in me a desire for food, and the fact that I believe this dish of Peking Duck will fulfill my desire seems to be a perfectly sound explanation of my eating it.

Even some anti-Humeans concede that desires are *occasionally* basic sources of motivation. For instance, we saw (in section 3.2.3) that Nagel recognizes a distinction between motivated and unmotivated desires. Motivated desires are desires that result from practical reasoning, but unmotivated desires “simply assail us” in the sense that they do not arise from any process of practical deliberation. Instead, unmotivated desires are *simply there* as the basic elements of an agent’s motivational set. They are the grounds on

\(^{19}\) Within the philosophy of mind there has been some skepticism as to whether folk psychology has any place in a mature neuroscience. The classic exponent of this sort of skepticism is Paul Churchland (1981). On the other hand, two prominent defenses of folk psychology’s explanatory success and its scientific credentials are Horgan & Woodward (1985) and Jackson and Pettit (1990).

\(^{20}\) A similar point is noted by Russ Shafer-Landau (2003, p. 132).
which new desires may be produced through *procedural* practical reasoning. For Nagel, the paradigmatic cases of unmotivated desires are the appetites, like hunger, and the emotions. Nagel seems to concede to the Humeans that desires for food are generated by procedural practical reasoning from one’s desire to eliminate one’s own hunger. My desire for this Peking Duck is prompted by my hunger, and my knowledge that eating the dish will sate my hunger. By making room for unmotivated desires in his theory of action, Nagel acknowledges that at least some human behaviors can be explained through the Humean belief-desire framework. The same can be said for intentional behavior in the other animals. When a rat presses a level which dispenses food pellets, the natural folk-psychological explanation is that it wants the pellets and believes lever-pressing is a means to retrieving them.

According to Morgan’s Canon, a (“higher”) psychological mechanism should be attributed to an agent only if it can explain behaviors that an alternative (“lower”) mechanism could not. If we follow the Canon, and if some behaviors in humans and nonhumans can adequately be captured by the Humean belief-desire model, then the burden is on the anti-Humean to show what, if any, behaviors can *only* be explained by anti-Humean mechanisms such as besires or desire-entailing beliefs. If there are no such behaviors, then Morgan’s Canon would call for dispensing with the anti-Humean mechanisms. The usual anti-Humean response to this challenge is to argue that there is no adequate way of explaining *moral* action other than by appeal to besires or desire-entailing beliefs. Thus, the debate concerns whether moral action *must* be motivated by
something other than desire plus means-ends belief. I believe that greater weight of
evidence suggests that motivational Humeanism can fully account for moral action.

Human beings are capable of numerous behaviors which are unseen in other
animals. One uniquely human behavior which many take to be morally significant is that
we are capable of engaging in directed altruism towards strangers with whom we have
neither genetic kinship nor social familiarity. By contrast, apes and other large-brained
animals seem to limit their altruistic behavior to kin and social companions (Fehr &
Fischbacher, 2003, p. 785; Silk, 2005). But humans extend their helpful actions to a much
broader range of recipients. We invest time and effort in helping the needy through
volunteer work. We make frequent anonymous donations to charity. We aid each other in
natural disasters. We even risk our lives for the sake of complete strangers in war or
times of crisis (Van Vugt & Lange, 2006, p. 237). Moreover, these altruistic acts are seen
as issuing from an ethical commitment to do what is right (Singer, 1995). They are
examples of moral conduct. Anti-Humeans can really revel in this human propensity to
help strangers. For it may turn out that the Humean belief-desire model lacks the
resources to explain it. If so, anti-Humean mechanisms may have to be introduced, and in
that case the appeal to besires, desire-entailing beliefs, or substantive practical reasoning
would not run afoul of Morgan’s Canon. To assess this potential reply to the argument
from Morgan’s Canon, let’s now consider in detail one argument designed to show that
Humeanism cannot explain human altruism toward strangers.
Stephen Darwall (1983) devised an interesting thought experiment in favor of anti-Humean moral motivation. In his view, “a person’s motivational capacities, in the broadest sense, are not constituted simply by his desires but also by capacities of imagination, sensitivity, and so on” (Darwall, 1983, p. 39). To support this anti-Humean claim, Darwall asks us to imagine Roberta—a woman who, for the first time, sees a film about the plight of textile workers in the United States. Roberta is “shocked and dismayed by the suffering she sees” (Darwall, 1983, pp. 39 – 40). After the film, Roberta attends a discussion about how to alleviate the situation, and she then makes a moral decision to participate in a boycott of a company that has been particularly abusive towards its workers. Darwall wants to attribute Roberta’s decision to join the boycott to “her becoming aware, in a particularly vivid way, of considerations that motivate her desire and that she takes as reasons for her decision: the unjustifiable suffering of the workers.” As we saw, the anti-Humean has several resources for characterizing Roberta’s shift in motivation. After perceiving the requirements presented by the situation of the textile workers, consistent with a moral conception of how to live, Roberta may have formed a besire or desire-entailing belief through which she simultaneously recognizes the injustice of the textile workers’ plight, and becomes motivated to join the boycott.

But Darwall denies that Roberta’s newfound ethical commitment must be explained, in a Humean spirit, by a general desire to relieve suffering, which Roberta already had before seeing the film, together with the means-ends belief that joining the boycott is a means to satisfying her desire. Instead, Darwall suggests that Roberta “may
have had no desire prior to viewing the film that explains her decision to join the boycott.” Darwall gives the following reason for ruling out this Humean hypothesis:

For [Roberta] to come to a general *desire* to relieve suffering she would have had to become actively concerned about the fact of suffering itself, conceived independently of who suffers, and she may never have done that (Darwall, 1983, p. 40).

According to Darwall, the Humean must ascribe to Roberta a general desire to relieve suffering which is abstracted from any particular individual. In that case, the content of this desire would be a state of affairs in which suffering is alleviated wherever and in whomever it occurs. The Humean may be tempted to characterize this desire to relieve suffering-in-general as either an other-regarding desire to ease the distress of the textile workers, or as a moral desire to change a state of affairs (viz., the workers’ suffering) which Roberta judges to be morally objectionable. Darwall, however, contests this hypothesis. He suggests that Roberta may not have any such desire to relieve suffering-in-general, because she may never have “become actively concerned about the fact of suffering itself.” He goes on to write that having a desire involves “dispositions to *think* about its object, to *inquire* into whether there are conditions that enable its realization” (Ibid.). And it’s conceivable that Roberta never reflected about suffering-in-general, nor about the conditions under which it occurs. Furthermore, Darwall invokes Hume’s distinction between sympathy and the love of mankind. For Hume, man is capable of being affected by the feelings of others, but from this fact it does not follow that we have a *general* desire for human well-being (Darwall, 1983, p. 41). Our capacity to sympathize with another person depends on the particular qualities of that person, and on the
relationship that person bears to oneself. So if we follow Darwall, we can be skeptical about whether there really is a desire, rather than some anti-Humean motivation, that could explain Roberta’s decision to help end the suffering of distant strangers. Roberta’s action, then, may be submitted as an instance of a behavior that can be explained only by anti-Humean psychological mechanisms. In that case, Morgan’s Canon would support postulating these mechanisms as the wellsprings of moral motivation.

However, even granting that Roberta is not motivated by a desire to alleviate suffering-in-general, I do not think that her action is beyond the scope of explanation in terms of the Humean belief-desire model. The Humean has a rich supply of theoretical resources to explain Roberta’s decision to join the boycott. What would be essential is that Roberta has some capacity to empathize with the textile workers. As we saw in the discussion of de Waal’s work, there are several levels of empathy. The most basic level is emotional contagion, whereby the subject’s emotional state matches the perceived emotional state of the other. This can be said to happen to Roberta if, upon seeing the film, she witnesses the suffering of the textile workers and comes to feel distress herself. Once the agent begins to share in the emotional state of the other, this emotion can induce one of several types of desire to help the other. Self-regarding desires concern some state of the agent him or herself. If Roberta feels vicarious distress for the textile workers, she may form a desire to end their suffering, if only as a means to end the distress she herself feels as a result of seeing the film. However, if this were Roberta’s motivation, the object of her desire would be to alleviate her own condition, and not that of the textile workers.
On the other hand, the object of the Roberta’s desire may be the textile workers themselves. In this case, Roberta’s vicarious distress would induce an other-regarding desire to relieve the suffering of the workers. According to de Waal’s analysis of empathy, if Roberta acts on an other-regarding desire to help the workers, her motivation would be classified as sympathetic concern. This is the second level of empathetic motivation in de Waal’s theory. Given the details of Darwall’s story, Roberta would also have to reach the third level of empathy—empathetic perspective-taking. After seeing the film, she joins a boycott of the goods produced by a company whose treatment of its workers was egregious. The implication is that Roberta sees joining the boycott as a means to improving the workers’ condition. In order to see the relevance of the boycott, Roberta would have needed to form a theory of the workers’ situation and their needs.

Contrary to what Darwall suggests, the self- or other-regarding desires that motivate Roberta to help the workers need not be a general desire to relieve suffering. Instead, I suggest that her desires could be aimed specifically at ameliorating the textile workers’ condition. Again, these desires could have been induced by Roberta’s emotional capacity for empathy. But her empathetic concern could have a specific target. Before ever seeing the film about the workers, Roberta may have had desires for the well-being of her own family and friends. Perhaps she had seen her family and friends experience similar hardships to the ones that she learns that the textile workers face. Perhaps, like the workers, Roberta’s loved ones struggled to make ends meet, or perhaps they’ve been blighted by poor health. Roberta might have wished desperately for the good of those she
cared for. These wishes may be seen as other-regarding desires whose objects Roberta was indeed disposed to think actively about. It’s furthermore conceivable that Roberta never would have noticed the suffering of the textile workers, were it not for seeing the film. Upon seeing it, though, she is reminded of the suffering that she witnessed her own family and friends endure. And assuming that Roberta has any empathetic inclinations in her, her emotions would be affected by the plight of the workers, just as they had been affected by that of her loved ones. Consequently, Roberta’s empathetic concern for the workers would induce an other-regarding desire to assist in efforts to improve their living conditions.

This story about the way one’s other-regarding desires might extend to include a broader range of individuals—even total strangers—is consistent with Hume’s account of the operation of sympathy. According to Hume, the tendency to share in the affections of others is more acute when the object of sympathy resembles the sympathizer, and when the sympathizer is nearer (i.e., contiguous) to the object of sympathy (Cohon, 2004). Thus Roberta comes to feel vicarious distress for the textile workers because she recognizes their suffering as similar to that experienced by people “closer to home,” viz. her loved ones. Current research on empathy bears out Hume’s claim that people empathize more with individuals who are more similar to them. Yet there is also substantial evidence that empathy promotes altruistic action toward strangers (Van Vugt & Lange, 2006, p. 249). I suggest that these two aspects of empathy are not in tension. Indeed, as I illustrated with Roberta, empathy towards strangers can grow out of empathy
towards familiars, provided that there are relevant similarities between strangers and familiars which enable the empathy one feels towards familiars to extend to others. Moreover, the process of *empathy-extension* that I just illustrated may be characterized as a process whereby an emotion induces an other-regarding desire to act in the interests of those with whom one empathizes.

Aside from an other-regarding desire, Roberta may have been motivated by a third type of desire to help the workers. This would be a *moral* desire to correct what Roberta judges to be an unjust, cruel, or otherwise immoral state of affairs. Again, the difference between a moral desire and a non-moral desire is that the former contains moral concepts. If Roberta is motivated by an other-regarding desire to help the workers, the content of her desire would refer to the needs and interests of the workers. Thus Roberta could describe her own other-regarding desire by saying “The workers need such-and-such and I want to help them get it.” On the other hand, if Roberta is motivated by a moral desire, the content of her desire would include moral concepts. Roberta could describe her own moral desire by saying “The workers’ condition is *unjust*, and I want to help change it.” Here the anti-Humean might insist that a moral desire can only arise from a belief, desire, or operation of substantive practical reasoning. However, the origins of moral desire can be explained without departing from the Humean claims that cognitive states are never sufficient to generate affective states, and that existing affective states are always necessary to produce new affective states.
In section 3.2.2, I suggested that desires can be induced by emotion. In my view, moral desires are induced by what Jonathan Haidt and other researchers call the *moral emotions* (cf. Haidt, 2003a; Moll et al., 2008). According to Haidt, moral emotions are “those emotions that are linked to the interests or welfare of society as a whole or at least of persons other than the judge or agent” (Haidt, 2003a). The moral emotions are grouped according to two characteristics: *elicitors* and *action-tendencies*. In his review of moral emotions research, Haidt says that moral emotions are generally elicited by situations and events that affect others (Haidt, 2003). But judging from the details of Haidt’s presentation, another critical factor which sets moral emotions apart from other emotions is their close link to moral judgment. For instance, *self-critical* moral emotions, such as guilt, shame, and embarrassment, are elicited when one judges oneself to have committed some kind of moral violation, while *other-critical* moral emotions, such as indignation/anger and contempt/disgust, are elicited when one judges someone else to have acted immorally (Haidt, 2003a; Moll et al., 2008, pp. 2 – 3, 6). Furthermore, moral emotions have effects on motivation. They’re said to generate *action-tendencies*—that is, they put a person in a motivational state in which “there is an increased tendency to engage in certain goal-related actions (e.g., revenge, affiliation, comforting, etc.)” (Haidt, 2003a). In my view, the action-tendencies of moral emotions provide evidence for the link between moral emotions and moral desires. Moral emotions have motivational effects by virtue of inducing moral desires in the agent.
Empathy (also characterized as “compassion”) is among the most prominent of the moral emotions. Roberta’s decision to help the textile workers could have been motivated by a moral desire induced by empathy. When a perceived moral violation leads to suffering, the observer may empathize with the victims. Anger is another moral emotion. It can be elicited by the obstruction of one’s goals, as well as episodes involving the morally-laden judgment of being betrayed, insulted, or treated unfairly. The “action tendencies” associated with anger include the motivation to redress situations the subject judges to be a moral violation. This motivation is present even when the subject him- or herself is only a third-party witness to events befalling others. In one of Haidt’s studies, participants were shown film clips portraying injustices, and were then asked to rate a variety of alternative endings. Participants reported being most satisfied by endings in which the perpetrator of injustice suffered (Haidt, 2003a).

Further evidence that anger motivates people to redress perceived moral wrongs can be drawn from an experimental economics paradigm involving the ultimatum game. In the ultimatum game, two players are shown a sum of money, say $10. The first player, called the “proposer,” is instructed to offer any whole number of dollars, from $1 to $10, to the second player, who is called the “responder.” The proposer can make only one offer, and the responder can either accept or reject this offer. If the responder accepts the offer, the money is shared according to the terms of the offer. If the responder rejects the offer, neither player receives anything. After the responder either accepts or rejects the offer, the game ends. Traditional game theory predicts that both players will maximize
their respective monetary payoffs. Thus the responder should accept any offer greater than 0, since receiving some money is better than nothing. The proposer, meanwhile, should anticipate the responder’s decision and make the smallest possible positive-value offer to the responder, thereby minimizing the amount he has to give up. However, when the experiment is run with people, this outcome is generally not observed (Gintis et al., 2007, p. 608). When the ultimatum game is played only once, under varying conditions and with varying amounts of money (including large sums of money), proposers usually offer 50% of the original sum and respondents reject offers below 20% about half of the time (Ibid.; Sanfey et al., 2003, p. 1755).

Participants in these experiments have reported that they reject low offers because of an angry reaction to an offer judged as unfair (Sanfey et al., 2003, p. 1756; cf. Pillutla & Murnighan, 1996). In addition, Sanfey et al. (2003) used fMRI scans on subjects playing the ultimatum game, and found that brain areas associated with anger were activated in responders by low or “unfair” offers of $2: $8 or less. Activation of the bilateral anterior insula is thought to be involved in the neural realization of specific other-critical moral emotions, including disgust and anger (Sanfey et al., 2003, p. 1757; Moll et al., 2008, p. 15). Greater activations of the bilateral anterior insula were correlated positively with the degree of an offer’s unfairness. The anterior insula exhibited greater activations in response to offers of $1: $9 than to $2: $8, and offers of $2: $8 generated greater activations than fair offers of $5: $5. Moreover, the magnitude of activation in this brain region correlated with subsequent decisions to reject the offer.
Participants with stronger anterior insula activations in response to unfair offers were more likely to reject an unfair offer. Also, the anterior insula exhibited more intense activation in response to unfair offers from a human proposer than unfair offers from a computer. From these results, Sanfey and colleagues conclude that activation of the anterior insula is involved in the neural realization of anger directed at persons who intentionally violate a principle of fairness. This in turn motivates resistance to unfairness in the form of a rejected offer (Sanfey et al., 2003, p. 1756). Based on these considerations I suggest that another way to explain why Roberta joins the boycott is that she is angered by the injustice inherent in the workers’ condition. In that case her anger would induce a moral desire to redress the injustice by joining the boycott.

An alternative hypothesis might save anti-Humeanism from this argument. It could be that moral emotions, and hence moral desires, are themselves explained by anti-Humean psychological mechanisms. Couldn’t we attribute the motivational force of a moral emotion or desire to the recognition of a reason to redress a moral wrong? This recognition might be achieved through a cognitive state, such as a desire-entailing belief, or by a state of mind with both cognitive and affective aspects, such as a desire. Thus, the responders in the ultimatum game are angered by offers of less than 20% of the reward because they have the desire or desire-entailing belief that such offers are unfair. But if this anti-Humean hypothesis were true, then ultimatum game responders who make the decision to reject an unfair offer should exhibit activity in brain areas associated with cognition. The anti-Humean might be encouraged by the fact that, in Sanfey’s
experiments, the dorsolateral prefrontal cortex (DLPFC) showed greater activity in response to unfair offers than fair offers from human partners. The DLPFC has been linked to cognitive processes, such as goal maintenance and executive control. However, even the DLPFC’s activity still does not support anti-Humeanism. For if the DLPFC were the cognitive structure primarily responsible for motivating the responder to reject an unfair offer, we’d expect to see greater activation of the DLPFC in cases where the responder rejects an unfair offer than when he accepts. But Sanfey and colleagues found that DLPFC activity is relatively constant across unfair offers, whether accepted or rejected, and that it was uncorrelated with respondents’ acceptance rates (Sanfey et al., 2003, p. 1757). Most likely, then, it’s not some type of cognitive state or hybrid state which detects a moral violation and then prompts an emotion or desire to reject an offer. The fact that the DLPFC was more active in response to unfair offers could be explained in one of two ways compatible with cognitive or non-cognitive Humeanism, respectively. It could indicate the presence of a moral belief (a cognitive state) that the offer was unfair, or it could indicate the additional cognitive demands imposed by the decision whether or not to accept an unfair offer. These two interpretations are not mutually exclusive, but Sanfey and colleagues favor the latter. They suggest that the heightened DLPFC activity in response to unfair offers indicates the subject’s experience of the conflict between receiving the benefit of a portion of the money reward and the emotional cost of letting an unfair offer go unpunished (Sanfey et al., 2003, p. 1757).
Darwall’s example of Roberta is intended to show that the explanation for a person’s moral choice to help distant strangers cannot be reduced to a Humean belief-desire explanation. However, in rebutting Darwall I’ve argued that cases of moral action can be explained by the Humean belief-desire model. More precisely, moral motivation can stem from reasoning about the means to satisfying emotion-induced desires. The case of Roberta is rather like that of the piping plover. By Morgan’s Canon, we should not attribute to the plover the capacity to represent the intentional states of other agents, because the plover does not produce the behaviors that only second-order intentionality would be able to produce. All its behaviors are within the explanatory scope of first-order intentionality. Likewise, I suggest, we should not attribute to moral agents besires, desire-entailing beliefs, or motivationally efficacious substantive reasoning, because moral agents do not exhibit behavior that only this anti-Humean machinery would generate.

3.5 The default status of Humeanism

Now I discuss a potential objection to both arguments adduced above. Shafer-Landau raises a similar worry (see Shafer-Landau, 2003, pp. 132 – 33). As the objection goes, the arguments from continuity and Morgan’s Canon unjustifiably ascribe a default status to Humeanism. In section 3.3, I drew on de Waal’s research to suggest that proto-moral behavior in apes and other social animals can be adequately explained by the Humean belief-desire model. Then, I suggested in section 3.4 that Humeanism can fully explain some non-moral behaviors in humans, like the pursuit of food when one is hungry. Because Humeanism can be extended to account for moral action, Morgan’s
Canon tells against postulating additional anti-Humean mechanisms to explain moral conduct. However, some philosophers are ready to argue that anti-Humean mechanisms can provide superior explanations even of non-moral behavior, at least in humans. Thomas Scanlon (1998), for instance, has disputed the Humean understanding of desire. On Michael Smith’s Humean account of desire, a desire is merely a mental representation of a state of affairs plus a disposition to bring it about. But Scanlon insists that having a desire cannot be separated from a tendency to judge that something is a reason for acting (Scanlon, 1998, p. 40). He is not as willing as Nagel to concede the existence of “unmotivated” desires which constitute ultimate sources of motivation. According to Scanlon, even the most primitive appetites do not simply assail us; rather, their motivational force is derived from judgments about one’s reasons for action. As a case in point, Scanlon offers an analysis of thirst (Scanlon, 1998, p. 38):

Suppose I am thirsty. What does this involve? First, there is the unpleasant sensation of dryness in my mouth and throat. Also, there is the thought that a cool drink would relieve this sensation and, in general, feel good…The present dryness in my throat, and the fact that this condition is not about to go away on its own, give me reason to believe that a drink of water in the near future will give this particular pleasure. But the motivational work seems to be done by my taking this future pleasure to count in favor of drinking.

As Scanlon would have it, the motivation to drink is generated ultimately by a tendency to see a reason for quenching one’s thirst, and not by a brute urge or disposition to eliminate unpleasant sensations.

Scanlon wants to argue that “having what is generally called a desire involves having a tendency to see something as a reason” (Scanlon, 1998, p. 39). He suggests that
a necessary condition for having a desire is this evaluative component of judging something to be a reason for action. For instance, Scanlon alludes to Warren Quinn’s example of a man who feels compelled to turn on every radio he sees. However, the man does not see anything good about turning on radios; he merely has an urge that compels him to do it. But precisely because the evaluative component is absent from this man’s motivation, Scanlon thinks it’s wrong to say he has a desire to turn on radios. On the other hand, Scanlon suggests that the Humean is committed to saying that this man is in a desire state. For on Michael Smith’s Humean characterization of desire, a desire is any intentional state with a world-to-mind direction of fit. This man’s urge to turn on radios satisfies that description. He has a representation of a state of affairs in which all radios are turned on, and he is motivated to realize that state. However, Scanlon’s point is that we often see no reason to do things which we feel an urge to do. In those cases, we are not motivated by genuine desires.

Scanlon supports his claims by considering the phenomenology of judgment and motivation (Scanlon, 1998, p. 34). In my view, such considerations have limited import. The theory of motivation is the science of what causes behavior. While it is desirable for a scientific explanation of behavior to cohere with evidence from introspection, we should also prefer a theory which successfully predicts observable behavior. I think Humeanism fares as well or better than anti-Humeanism on both counts. In particular, a Humean framework can accommodate Scanlon’s phenomenological points regarding the

21 For a defense of this claim, refer to chapter 2, section 2.4.
seemingly tight link between having a desire and the experience of seeing things as reasons. Even on a Humean theory of desire, for instance, the radio man may not have a desire to turn on all radios. Jesse Prinz contends that the term “desire” carries multiple meanings. Among these meanings, desire may be understood as Smith understands them—a disposition which causes one to behave in predictable ways. However, desire may also be understood as an affective state. An affective state, according to Prinz, is a state which contains a valence marker, and a valence marker is an internal reinforcer of behavior. Valence markers have positive or negative value, depending on their typical effects on behavior. Positive reinforcers increase the probability that a behavior will be continued, whereas negative reinforcers increase the probability that a behavior will be terminated (Prinz, 2004, pp. 174 – 73). From a phenomenological point of view, valence markers can be experienced as emotions. Joy is a positive valence marker, which makes it more likely for one to preserve the object of one’s joy; anger is a negative valence marker, which makes it more likely for one to destroy the object of one’s anger. Valence markers can also be experienced as states of pleasure or pain.

Prinz suggests that it’s possible for behavior not to be caused by an affective state. But in those cases it does not seem right to say that such behavior is caused by desire. Prinz cites an experiment conducted by Kent Berridge where the “affective component” of hunger is separated from its “action-driving component” (cf. Prinz, 2004, p. 195; Berridge, 1996). In one treatment, Berridge stimulated the parts of rats’ brains believed to

---

22 For a discussion of these various meanings, see pp. 196 – 97 of Prinz (2004).
induce eating behavior. Rats in this condition would gorge themselves on all foods they were given—even ones they normally disliked. Meanwhile, these rats displayed aversive behaviors at every bite (face-wiping, head-shaking, forelimb-flailing, etc.). As Prinz sees it, something compels Berridge’s rats to eat, but they do not have a desire to eat because their eating behavior is not motivated by the affective state of taking pleasure in what they eat. Prinz also notes that a comparison can be drawn between Berridge’s rats and human drug addicts. Addicts often pursue their favored drug even after it ceases to bring pleasure. But if so, the addict’s pursuit would not be motivated by a desire.

The lesson to be taken from Prinz is that a Humean need not reduce desire to an intentional state plus a behavioral disposition. On at least some interpretations of the concept of desire, an affective component may be essential. If the radio man takes no pleasure in turning on every radio he sees, or if in general his behavior is not caused by any affective state at all, even a Humean can recognize that this man has no desire to turn on radios. Furthermore, the Humean can agree with Scanlon in holding that a desire must involve judgments about one’s reasons for action. Scanlon tells us that we’re motivated to quench our thirst because we expect that the sensation of it will feel good, and we judge that this pleasure gives us a reason to drink something. But on this point there may not be any disagreement between Scanlon and the Humean. The Humean theory of motivation (HTM* above) allows for a creature to be motivated by instrumental reasons for action—i.e., reasons to act based on a justified belief that some action is a means to a desired end (cf. chapter 1, section 1.1, as well as (H1*) and (H2*) above). If one desires the pleasure
that comes with quenching one’s thirst, and one believes that by drinking something, one will do it, then this would constitute an instrumental reason to drink. So even for the Humean, judgments about reasons can motivate. But these motivationally effective judgments will be judgments about instrumental reasons.

However, Scanlon insists that seeing a reason to quench one’s thirst is not a judgment about an instrumental reason. He thinks instead that judgments about reasons can be basic sources of motivation which do not require the presence of any pre-existing desire or other motivational state (cf. especially Scanlon, 1998, pp. 33 – 34). Scanlon’s point in discussing the example of water seems to be this: to judge the experience of enjoying a cool drink in a favorable light is to recognize a reason to drink something cool. Recognizing a reason need not be experienced as recognizing a desire plus a means-ends belief. Still, such recognition may be sufficient to motivate drinking something cool. Scanlon’s phenomenological analysis of desire may seem perfectly sensible. Perhaps most laypeople without a philosophical ax to grind would be satisfied by it. Perhaps Humeanism is not so central to the folk theory of motivation as to render unacceptable any reasons explanation of an action which does not cite a desire plus a means-ends belief as the explanans. So, Scanlon might continue, we should take his phenomenological analysis at face value and conclude that there are non-instrumental reasons to fulfill even our most primitive appetites, and these reasons are sufficient to explain action.
But even when it comes to phenomenology, the least that can be said is that Humeanism is at no disadvantage to Scanlon’s view. It certainly wouldn’t sound wrong to characterize one’s reason to drink something cool as an instrumental reason, especially once we’ve adjusted the concept of desire to include an experience of affect. The phenomenological considerations adduced by Scanlon do not identify a decisive advantage for anti-Humeanism. On the other hand, there are non-phenomenological considerations which suggest in favor of taking Humeanism as the default approach to motivation. In section 3.3, I argued that Humeanism has the advantage of explaining the “proto-moral” behavior of non-human animals. For anti-Humean motivations like besires and desire-entailing beliefs seem to lie beyond the cognitive reach of non-humans. Animals presumably cannot engage in substantive practical reasoning about objective values, or form a conception of how to live. Considerations of theoretical continuity suggest in favor of postulating human psychological capacities which are evolutionary modifications of non-human capacities. Thus we should expect that the mechanisms which explain human moral motivation are (probably much more complex) modifications of those which explain proto-moral behavior in animals. Since proto-moral behavior can be explained by a Humean belief-desire model, we should at least see traces of the Humean model in the best account of moral motivation. In section 3.4, I noted that it is common practice to explain a broad range of non-moral behaviors, like the pursuit of food or water, by way of the belief-desire model. When it comes to behaviors like these, Humeanism has in fact been the default theory. Perhaps our explanatory habits are deeply
mistaken, but such a mistake would be exposed only if there are facts which Humeanism cannot account for. In the last section and in the present one, I’ve endeavored to show that nothing which has been alleged by anti-Humean philosophers to fall beyond the explanatory scope of Humeanism actually does so. When room is made for moral emotions and moral desires, the Humean belief-desire model can be extended to explain moral behavior.

The default status of Humeanism is also supported by evidence from studies in neurobiology and neureconomics which attempt to isolate the cerebral mechanisms responsible for moral action. For instance, a large body of evidence shows that, in animal brains, the ventral striatum and nucleus accumbens, along with the insula and the orbitofrontal cortex, respond to the satisfaction of basic biological needs such as food, shelter, and social contact. In humans, these same areas also respond to receiving money. But one study by Moll et al. (2006) reports that increased activity in the ventral striatum can be observed with fMRI scanning when human subjects choose to donate money to a charity (cited in Mayr, Harbaugh, & Tankersley, 2009, p. 310). In another study by Harbaugh et al. (2007; also cited in Mayr, Harbaugh, & Tankersley, 2009, p. 310), researchers compared human subjects’ neural reactions to receiving money for themselves and being forced to give it away to a charity. In the first treatment, the researchers forced subjects to donate money to a charity in a tax-like condition, and measured ensuing increases in neural activity in the ventral striatum, the head of the caudate, and the nucleus accumbens. In a second treatment, they measured increased
activity in the same areas when subjects were given a sum of money. The authors discovered that the difference between these two measures reliably predicts people’s willingness to donate to charity on a *voluntary* basis. People who exhibited larger neural responses to making mandatory donations, relative to receiving money, were more willing to donate voluntarily (Ibid., p. 311).

Harbaugh et al. suggest that activity in the ventral striatum and other brain regions is a neural indicator of the *utility* a subject receives both from getting money and from giving it to charity. Here, “utility” is a technical term for the satisfaction of desire. The authors support this conjecture by using increases in neural activity in the ventral striatum and insula measured during the two treatments as a basis for plotting *indifference curves* for each subject. In microeconomic theory, indifference curves are used to represent a consumer’s preferences among alternative bundles of at least two goods. For instance, in the Harbaugh et al. study, indifference curves were used to represent all combinations of money received relative to money donated which yield the same level of utility for the subject. The subject’s preferences are determined by his or her *marginal utility* for money received versus money donated. The subject’s marginal utility for getting money is a measure of the additional satisfaction he or she obtains from an additional unit of money received, and similarly the subject’s marginal utility for donating money to charity measures the additional satisfaction he or she obtains from an additional unit of money donated. Harbaugh and colleagues used increases in neural activity exhibited by each subject in response to giving and receiving money as estimates for marginal utilities. The
ratio of the two marginal utilities determines the subject’s *marginal rate of substitution* of money received for money donated. Conceptually, the marginal rate of substitution represents a person’s willingness to give up one good in exchange for more of another good. Graphically, the marginal rate of substitution is the slope of the person’s indifference curves for any given combination of money received relative to money donated. A selfish person who strongly prefers getting money to donating money would only be willing to give up very little of his money in order to donate an additional unit of it. But an altruistic person who has a stronger preference for donating money would be willing to give up considerably more of his money in order to donate.

Harbaugh and colleagues found that by modeling their subjects as utility maximizers in this way, they could predict how much they were willing to donate voluntarily to a charity. As I emphasized in chapter 1 (section 1.1), the utility-maximizing model which lies at the core of rational choice theory is itself a version of motivational Humeanism. Utility-maximizing agents are instrumentally rational agents who act to maximize the satisfaction of their desires. Harbaugh’s findings provide compelling empirical evidence in favor of Humeanism, because they show how a belief-desire model can be used to successfully predict moral action. To my knowledge, anti-Humeans have not attempted to subject their theories to such rigorous empirical test. The relative predictive and explanatory success of Humeanism is yet another reason to accept it as the default theory of motivation, both moral and non-moral.
3.6 Conclusion

This completes my case for the presumptive superiority of the Humean theory of moral motivation. Considerations of theoretical parsimony and continuity throw the burden of proof onto the anti-Humeans, who must not only identify moral behaviors that could be explained only by their mechanisms of choice, but also provide an account of the origins of these mechanisms which is compatible with evolutionary theory. Humeans have come a long way in meeting both these tasks. Anti-Humeans have not. And until they do, Humeanism will continue to have the upper hand. Furthermore, insofar as Humeanism holds that moral conduct is motivated by instrumental reasons for action, instrumentalism offers the best going account of reasons to be moral. Because it outperforms its rivals in showing how moral reasons can come to motivate, it best satisfies Williams’s explanatory constraint on practical reasons.

There are, however, important challenges to instrumentalism that still have to be addressed. According to instrumentalism, combinations of desires and means-ends beliefs constitute reasons for action. But critics of instrumentalism have expressed doubt that desires can ever play a role in justifying action. Another crucial source of resistance to instrumentalism is encapsulated by the Ted Bundy Paradox (section 3.1). If all reasons to be moral are instrumental reasons, the fact that people have wildly different desires would make moral reasons contingent in a way that they are not. Persons with violent and sadistic desires might not have any reasons to act as morality requires. Instead, depending
on what desires they happen to have, some people’s best reasons may favor acting immorally. In the next two chapters, I try to lay these remaining worries to rest.
4. Chapter 4: The practical evaluation of ends

4.0 Introduction

Instrumentalism is the view that all practical reasons for action are instrumental reasons. An instrumental reason, recall, is a normative practical reason to act in a way that optimally satisfies one’s desires (cf. chapter 1, section 1.1). For the instrumentalist, the sole criterion of practical rationality is the Instrumentalist Dictum. This principle tells us that our best reason for action is to do whatever maximizes the net satisfaction of our desires. However, a major criticism of instrumentalism is that it imposes unacceptably weak constraints on the desires that an agent may rationally act upon. The charge is that the Instrumentalist Dictum does not provide a basis for evaluating the desires which would be satisfied by instrumentally rational action. If, to take Warren Quinn’s example, a man would maximize the satisfaction of his desires by turning on every radio he sees, then his best reason for action is to do just that. And if he would maximize the satisfaction of his desires by murdering innocent people, then murder is the action favored by his best reason. Yet there is something compelling about the idea that people’s desires can be rational or irrational in and of themselves. It seems that the desires to murder innocent people or turn on all radios are not rational desires for anyone to have. Instrumentalism, however, lacks the conceptual resources to show that these desires are irrational. This problem motivates the most challenging objection to instrumentalism: even if an agent acts to maximize the satisfaction of her desires, she may nevertheless fail to be practically rational due to the irrationality of her desires. If this counterproposal is
true, then reasons externalism and counterfactual internalism have a distinct advantage over instrumentalism, for they offer criteria by which to evaluate the practical rationality of desire.

In this chapter, I respond to the charge that instrumentalism cannot provide a basis for criticizing desires. I argue, to the contrary, that instrumentalism does establish criteria for the normative evaluation of desired ends. On the other hand, I concede that instrumentalism permits the conceptual possibility that an instrumentally rational agent could have bizarre, self-destructive, or immoral desires. But in the face of this concession, I do not flinch. Instead I propose that, given certain empirical facts about what human beings typically desire, it is improbable that an agent with these intuitively faulty desires will turn out to be instrumentally rational.

In section 4.1, I consider several of Thomas Scanlon’s attempts to argue that desires never provide reasons for action. I suggest Scanlon systematically overlooks the fact that desires do seem to have a reason-giving role in ordinary discourse about practical rationality. I also suggest that instrumentalism provides a powerful explanation for the intuition that practical reasons can sometimes vary from person to person. I call this intuition the subjectivity of practical reasons. Next, in section 4.2, I emphasize a seemingly antithetical intuition which I call the practical fallibility of ends. According to this intuition, there are some ends which cannot be desired or pursued by any rational agent. I note the charge that instrumentalism cannot accommodate the fallibility of ends, and I discuss how theories of external reasons for action seem to provide an attractive
account of it. In section 4.3, I explore Philippa Foot’s neo-Aristotelian theory of external reasons. I reject this theory on the grounds that it’s incompatible with the subjectivity of practical reasons. Finally, in section 4.4, I present an exposition of David Schmidtz’s *reflective instrumentalism*. This more sophisticated version of instrumentalism purports to establish a normative foundation for evaluating all of an agent’s desired ends, but it continues to hold that agents can *only* have instrumental reasons in favor of choosing or pursuing ends (as defined in chapter 1, section 1.1). I offer reflective instrumentalism as the account of practical rationality which most successfully reconciles the subjectivity of reasons and the fallibility of ends.

### 4.1 Can desires ground practical reasons?

Thomas Scanlon attacks the claim that “reasons for action can or must derive their justificatory force from the agent’s desires” (Scanlon, 1998, p. 41). Instrumentalists are committed to this claim, since according to instrumentalism all practical reasons are instrumental reasons, and instrumental reasons are by definition considerations that favor acting in a way that maximally satisfies one’s desires. Scanlon also defends the contrary thesis that “desires almost never provide reasons for action in the way described by the standard desire model” (Scanlon, 1998, p. 43). The “standard desire model” which Scanlon describes is, in its essential features, instrumentalism about reasons. Scanlon then presents a series of thought experiments designed to show that merely having a desire does not entail a reason for action. In this section, I wish to show that instrumentalism withstands Scanlon’s criticisms. More than that, I wish to push back on
Scanlon’s arguments. I suggest that his account of practical rationality cannot adequately explain why different people can have different reasons for action. Instrumentalism, by contrast, provides an attractive explanation for this.

Scanlon has a preferred conception of desire which he calls “desire in the directed-attention sense” (Scanlon, 1998, p. 39). A desire in this sense is not a mere behavioral disposition to bring about a certain state of affairs. Rather, as Scanlon defines it (Ibid.):

A person has a desire in the directed-attention sense that P if the thought of P keeps occurring to him or her in a favorable light, that is to say, if the person’s attention is directed insistently toward considerations that present themselves as counting in favor of P.

In short, having a desire for P in the directed-attention sense is having a tendency to judge certain considerations about P as reasons for action. Although Scanlon thinks the Humean conception of desire lacks this evaluative component of “seeing something as a reason,” I argued in chapter 3 (section 3.5) that this evaluative aspect can be included within a Humean phenomenology of desire. A Humean would only specify that “seeing something as a reason” consists in seeing something as an instrumental reason—i.e., as a means to satisfying a desired end.

Let us hereafter adopt Scanlon’s notion of desire. Even if we do this, Scanlon argues that the considerations which present themselves as reasons via one’s desires-in-the-directed-attention-sense may not be reasons in fact. For a desire in the directed-attention sense can dispose an agent to make incorrect judgments about his or her reasons. To illustrate, Scanlon imagines a case, involving himself, which I call computer
Scanlon is beset by desire to purchase a new computer. He finds himself looking at computer advertisements in the newspaper, and thinking about how the features of various new models count in favor of buying one. In this scenario, Scanlon has a desire in the directed-attention sense to buy a new computer, since he is in a state that involves a tendency to *judge* that he has a reason to buy a new computer. However, Scanlon denies that his having this desire is sufficient to give him a reason to buy a computer. For his judgments regarding his reasons may be incorrect. He may, on further reflection, conclude that the features of the new models are of no real benefit. In Scanlon’s analysis, it’s not that his desire presents a reason to buy a new computer which is “outweighed” by other considerations. Rather, the desire provides no reason at all (Scanlon, 1998, p. 44).

On the other hand, suppose that Scanlon, after further reflection, *does* decide that he has a sufficient reason to buy a new computer. According to Scanlon, this will not be due to the fact that he desires a computer. Rather, his reasons for buying himself a new computer will be founded in considerations such as the fact that he would enjoy having it, that it would help him with his work, that it would impress his friends, and so on (Ibid.).

The case of computer lust does not support Scanlon’s view as strongly as he might like. The thought experiment has the feel of an *illustration* of his thesis, rather than an argument for it. Particularly unconvincing is Scanlon’s assertion that his desire to buy a computer does not present him with a reason. One wonders, why not? The story could very plausibly be dissected in a way that’s consistent with instrumentalism. Scanlon’s desire to buy the fancy computer he lusts after may still provide *some* reason for action,
although it may not be his *best* reason. The fact that he desires a new computer may still constitute a reason to buy one, but acting on this reason may preclude the satisfaction of other, stronger desires that Scanlon might have. It may turn out that he would maximize the satisfaction of his desires by spending his money on other things, provided that his desires for these other goods are stronger than his computer lust. In that case, Scanlon’s best instrumental reason may be to spend his money differently.¹ If asked to justify his decision not to buy a new computer, it seems perfectly sensible for Scanlon to give a rationale like the following: “As much as I wanted to buy a brand new computer, I realized that I could spend my money on things I wanted even more. So the reasons in favor of buying a computer were outweighed by reasons against it.” Scanlon seems to think that this is a distorted account of practical reasoning. In his view, the mere fact that he wants certain other things more than he wants to have a new computer does not in itself count as a reason against buying a new computer. Whether or not Scanlon has a reason to buy a new computer depends on the specific considerations which weigh in favor of or against buying one, and not on the desires he happens to have. Yet there is nothing obviously distorted about a rationale in which Scanlon justifies his decision not to buy a new computer simply by stating “There were other things I wanted more,” and leaving it at that.

¹ For a definition of a “better” instrumental reason, see the instrumentalist thesis (I-2) in chapter 1, section 1.1.
Scanlon says that he has *no* reason to buy a computer, despite having a desire to do so. But he fails to explain what exactly prevents him from having this reason. If the suggestion is simply that he cannot have a reason to buy a computer despite desiring one, Scanlon begs the question against instrumentalism. Indeed, when Scanlon segues into another thought experiment in which he supposes he *does* have reason to buy a new computer, he rules out by fiat the notion that his reason could be grounded in his desires. According to Scanlon, his reason for buying a computer is not that it will satisfy a desire, but that he will enjoy having it, that it will help him with his work, that it will impress his colleagues, etc. But why can’t these factors constitute reasons by virtue of being objects of *other* desires that Scanlon has? Suppose we demanded that Scanlon justify his decision to buy a new computer, and he gives the following reply: “First I thought about all the things I wanted in a new computer. I wanted one that I’d enjoy having, one that would help my work, impress my colleagues, and so on. Then, after studying various computer models, I surmised that owning model $X$ would best satisfy all these desires. So, I decided to buy model $X$.” Once again, this rationale seems perfectly sensible and appropriate. If there is anything defective about it, the defect isn’t obvious, and Scanlon does nothing to expose it (that is, in a non question-begging way). And yet this rationale implies exactly what Scanlon insists cannot be: that the mere fact of wanting a computer with certain features gives one a reason to own a computer with those features. If it’s true that practical reasons cannot be grounded in desires, Scanlon’s case of computer lust does not demonstrate why.
Scanlon’s other thought experiments are similarly inconclusive. None of them forbid a plausible re-analysis along instrumentalist lines. The same problem emerges in Scanlon’s analysis of the sensitivity of practical reasons to what he calls “subjective reactions.” Consider the following remarks (Scanlon, 1998, p. 42):

It is easy to accept the claim that my reasons for eating coffee ice cream and for going to the seashore rather than to the mountains depend on the fact that these things appeal to me. And this is true not only of reasons that are trivial or have to do with “matters of taste.” My reasons to help and support my friends and loved ones, for example, depend on the fact that they are my friends and loved ones, hence on my affection for them. A large part of the point of eating ice cream or taking a vacation is doing something that I will enjoy, so one’s “subjective reactions” are obviously of prime significance to the reasons one has for doing these things one way rather than another. And since affection lies at the core of relationships of love and friendship, there is nothing deflationary about the observation that these relationships depend in part on “how we feel” about the people in question (although it would be odd to say that the reasons these relationships involve depend on our “desires”).

Scanlon recognizes that, in some contexts at least, one’s reasons for action can be determined by one’s “subjective reactions” to things. But obviously, different people’s subjective reactions can vary wildly. It is because of these differences in subjective reactions that people can have divergent practical reasons. Hence the fact that people who enjoy eating coffee ice cream have a reason to eat it, whereas people who dislike the taste of coffee have no such reason. Further, a single person’s subjective reactions to things can fluctuate over the course of a lifetime. As a result of these changes, one can gain or lose reasons for action. A man’s amorous feelings give him reasons to treat his lover in certain ways. But if those feelings ever begin to subside, he would then have reasons to treat her differently. In sum, an agent’s reasons for action can, at least in some cases, be
contingent on her subjective reactions. Hereafter, I refer to this phenomenon as the 
subjectivity of practical reasons.

Instrumentalism can offer a powerful explanation for the subjectivity of reasons. The reason-giving “subjective reactions” described by Scanlon can be construed as desires. After all, desires vary from agent to agent. It’s a familiar fact that different people want different things. In addition, a given agent’s desires can change with time: today I might want to wake up early, but tomorrow I might prefer to sleep in. Let us call this phenomenon—the variability of desire from one agent to another, and from one time to the next—the contingency of desire. If reasons for action are based on desires, as instrumental reasons are, then it’s easy to see why practical reasons so often appear to be subjective. If the subjectivity of reasons is explained by the contingency of desire, this will support instrumentalism.

Moreover, if instrumentalism yields a better explanation for the subjectivity of reasons than competing views, that would make further evidence for instrumentalism. Scanlon, however, disputes the instrumentalist explanation. He resists characterizing a person’s reason-giving “subjective reactions” as desires. It is not, Scanlon tells us, one’s desires that provide reasons to act a certain way toward one’s lover, but rather one’s feelings toward him or her. But I think there are perfectly sensible ways of talking about the reasons we have to act toward friends and loved ones that do make reference to desires. Suppose Boy meets Girl, Boy falls madly in love with Girl, and comes to the conclusion that he has every reason to marry Girl. If we demand from Boy a justification
for proposing to Girl, the reasons he gives might include heartfelt declarations like “I want to share my life with her,” “I want to have children with her,” or “I never want to be without her.” It seems quite appropriate—even unremarkably so—for Boy to invoke these desires as elements in the practical justification of his decision to marry. Even if we were to be very persnickety, and demand from Boy an account of why these desires are relevant to his decision to marry, it would not sound infelicitous for him to offer the simple reply that “Marrying Girl would bring me these things that I want so dearly.”

In addition, the instrumentalist can account for why it’s so natural to say that Boy has a reason to marry Girl because of his feelings for her. Instrumentalists subscribe to the Humean theory of motivation. I explained in chapter 3 (section 3.2.2) that motivational Humeans argue in various ways for a regular connection between emotion and desire. I announced my allegiance to what Finn Spicer (2004) calls a nomological view of this connection. On the nomological view, emotions robustly co-occur with beliefs and desires, and this robust co-occurrence is explained by causal relations between emotions and these other states. Thus, emotions are capable of inducing combinations of beliefs and desires that can both motivate agents and provide them with instrumental reasons for action. For instance, Boy’s love for Girl may induce in him a strong desire to have and hold Girl forever as his wife. At the same time, his feelings of love may direct his imagination and his beliefs about the future in ways that strengthen his desire for the married life even further. An instrumentalist would identify this desire to marry as the basis for an instrumental reason to propose to Girl. But if the nomological
theory is true, then the instrumentalist can appreciate that Boy’s instrumental reason to propose derives from his feelings for Girl in the sense that his desire was produced by his feelings. Instrumentalism is thus not incompatible with the commonsensical notion Scanlon underscores—viz., that our reasons to act toward friends and loved ones are based on our feelings toward them. However, when it comes to giving a precise account of the relationship between emotions and reasons, the instrumentalist contends that emotions support practical reasons in virtue of having some regular connection (e.g., a causal connection) to desire. Evidence for this contention lies in the fact that people commonly and sensibly do cite their own desires—e.g., desires to grow old with someone, to have children with them, etc.—when rationalizing their actions toward friends and loved ones. There was, for instance, nothing counter-intuitive about the way that Boy rationalized his decision to marry Girl in the preceding paragraph.

Scanlon could easily reply that Boy’s declarations of love for Girl, framed in terms of his desires to be with Girl, do not really do the work of justifying his decision to marry Girl. For Boy may not have any reasons for his desires. To back up this objection, Scanlon might point out that people sometimes claim to fall in love for no reason. This is perhaps why aphorisms like Pedro Calderon de la Barca’s ring so true: “When love is not madness, it is not love.” However, the instrumentalist can accommodate this intuition. Love might induce a ground-level desire for a final end—i.e., an end pursued for its own sake rather than as a means to achieving another end (cf. chapter 1, section 1.1). For instance, Boy might not be able to articulate any normative reasons to justify why he
desires to share his life with Girl. At best, he might only be able to explain this desire by citing his amorous feelings for Girl. Sharing his life with Girl is something that Boy desires for its own sake. On an instrumentalist account, Boy’s desired end of sharing his life with Girl is not open to rational criticism, unless Boy has better instrumental reasons for doing something else with his life. This follows straightforwardly from the instrumentalist thesis (I-5), which I stated back in chapter 1 (section 1.1). So instrumentalism is consistent with the idea that people can have no reason to love others. But, if instrumentalism is true, it doesn’t follow that people who love without reason also lack reason to act on their feelings. I think this is an advantage of instrumentalism. For it can explain why people have particular reasons for acting in certain ways toward those they love: idiosyncratic feelings of love induce idiosyncratic desires, which support idiosyncratic instrumental reasons for the agent.

Of course, Scanlon can reply to these comments by repeating his basic point: a mere desire cannot support a reason for action. So, if Boy loves Girl for no reason, then in spite of all his desires to be with her, he has no reasons to act towards Girl. Scanlon deploys several more arguments to disprove the notion that the subjectivity of practical reasons is due to the contingency of desire. One of them is particularly challenging, though it also provides an opportunity for the instrumentalist to push back. Scanlon takes up Williams’s character of Owen Wingrave, which I broached in chapter 1 (section 1.4.3). As the story goes, the Wingraves are a family with a time-honored military tradition. Owen’s father urges his son to embrace the military life, just as his male
forbears had. But Owen himself is “left cold” by the prospect of a career in the military. Despite this, Owen’s father insists that he still has a reason to join the army “because of the inherent value in this calling and because family tradition requires it” (Scanlon, 1998, p. 48). Williams uses this example to support the main thesis of his actual reasons internalism—namely, that one’s reasons for action are derived from one’s existing motivations. In Williams’s analysis, Owen lacks any reason to join the army because he doesn’t have a motivation to do so. An instrumentalist would make a more specific claim than this: Owen lacks a reason to adopt the military life because he lacks a desire which would be best satisfied by joining the military. Thus, Owen Wingrave can count as a thought experiment which bears intuitive support for the claim that the subjectivity of reasons is explained by the agent’s desires.

Scanlon agrees that Owen should not pursue a lifestyle which “leaves him cold,” but not for the same reason as Williams and the instrumentalist suggest. It’s not that Owen shouldn’t become a military man because he doesn’t want to. Rather, Scanlon thinks he shouldn’t bend to his father’s wishes because

…one has good reason to choose as a career only one of those to which one is drawn—that is to say, which one has a desire in the directed-attention sense to pursue…being drawn to a pursuit is (at least under favorable conditions) a condition for having a good reason to undertake it as a career. For one thing, if one takes up a career which “leaves one cold,” then one is unlikely to succeed in it. For another, it’s not a good thing if every morning, in order to get yourself to go to work, you have to rehearse (or have someone else urge on you) all the reasons why it is worthwhile or honorable or dashing to be a military man… (Scanlon, 1998, p. 49).
Scanlon’s points here, while astute, distract attention from the relevant issue. Indeed, the considerations cited in the passage do seem to be powerful reasons for anyone to resist living their lives by others’ dictates. But this does not weaken the instrumentalist’s analysis of Owen’s case. It’s still quite plausible that Owen lacks a reason to join the army simply because he doesn’t want to do it. If so, then Scanlon is wrong to claim that desires can’t provide reasons. I will support this analysis by considering an adjustment of Williams’s thought experiment. It unfortunately involves sacrificing the story’s dramatic flare, but it will make my philosophical point more clear. Suppose Owen wasn’t such a black sheep, and the elder Wingrave had been much more successful at inculcating in his son a reverence for the family’s military tradition. Owen agrees wholeheartedly with his father that a military career would be a noble and distinguished way to lead his life, and the sense that he has a duty to carry on his family’s military tradition resonates strongly with him. But let’s also imagine that Wingrave Sr. isn’t quite so overbearing as Henry James makes him. He wants his boy to exercise some choice over his own life. So one day Owen’s father sits him down and speaks to him in earnest: “Owen, you and I both understand the importance of tradition to this family, and we both know that you would excel in the army just as all Wingraves did before you. That said, I want you to look into your heart and decide for yourself whether that kind of life is truly for you. Whatever the path you choose, I’ll still be proud to call you my son.”

Uninteresting as this after-school special adaptation of James’s tale may be, it has its philosophical advantages. Suppose Owen listens intently to his father’s words, and
takes some time to reflect on his future course. He considers several worthy occupations
to which he might devote himself. Finally, after much thought, Owen reaches the
conclusion that a career in medicine would be every bit as noble and laudable as military
service. Indeed, for every praiseworthy virtue or interesting aspect of a military
profession that Owen can think of, he can summon to mind an *equally* venerable feature
of the medical calling. Moreover, Owen correctly anticipates that a man of his extensive
talents would have no less aptitude for medicine than he would as a soldier, and, being
blessed with the Wingraves’ characteristic ambition, he would be no less driven to
succeed in one vocation than the other. Essentially, let’s assume that for every
consideration that favors Owen’s joining the military, there is an *equally* weighty
consideration in favor of his becoming a physician.

But now let’s admit just one difference: Owen has a stronger desire to be a
physician than a military man. In that case, it seems that Owen would have a sufficient
reason to become a doctor instead of a soldier. And, contrary to Scanlon’s insistence, he
would have this reason solely in virtue of his preference for a medical career over a
military one. Thus, when it’s assumed that all other considerations equally favor the two
alternatives in Owen’s dilemma, it seems that a desire which draws Owen more toward
medicine breaks the tie and provides Owen with a reason to become a doctor. Someone
else in Owen’s situation—say, one of his deceased forebears—may have chosen
differently. For the Wingrave ancestor, a desiderative tie between medicine and the
military might have been broken by a stronger desire to be a soldier. If so, his best
reasons would seem to favor a military life. Different people are drawn by their desires to take on different pursuits, and it does seem that these differences at least partially explain the subjectivity of practical reasons. Contrary to Scanlon’s suggestion, a person’s desires do provide him or her with reasons for action.

Scanlon could reply to this argument in one of two ways. First, he could deny that Owen’s bare preference for being a physician breaks the tie and makes pursuing a medical career his best reason for action. If, as Scanlon holds, merely having a desire for \( X \) cannot support a reason to \( X \), then Owen has no better reason to become a doctor than a military man. But I think the thought experiment reveals this reply to be unsatisfying. A more promising reply might be that Owen does have a better reason to pursue a medical career, but this reason is not based on Owen’s preference for becoming a doctor. Instead, it’s based on the fact that Owen will derive more enjoyment or satisfaction from doing that which he most wants to do. If “enjoyment” or “satisfaction” are supposed to be experiences that can be separated from the fulfillment of Owen’s desire to be a physician, we can simply adjust the thought experiment to ensure that Owen would equally “enjoy,” or be equally “satisfied” by, a military career as he would by a medical career. And if we maintain the assumption that Owen prefers to be a doctor, it still seems that Owen’s merely having a stronger desire to become a doctor breaks the tie, thereby giving him a better reason to go into medicine. On the other hand, it seems to me impossible to separate the experience of “enjoyment” or “satisfaction” from the fulfillment of Owen’s desire. In chapter 3, section 3.5, I argued that desire necessarily involves an affective

288
component. A desire is an affective state with positive or negative valence. Positive valence can be experienced as pleasure, or as emotions like joy or contentment. On my view, then, to say that Owen will derive greater “enjoyment” or “satisfaction” from a medical career than a military one is equivalent to saying that he has a stronger desire for the former. And if that’s correct, then Owen will have a better reason to go into medicine solely in virtue of his preference for a medical career.

The adjusted thought experiment with Owen Wingrave supports the thesis that desires can provide reasons for action. This thesis is the basic claim of instrumentalism. Scanlon denies this claim at the price of failing to account for the subjectivity of reasons. On the other hand, even if it’s true that an agent’s desires can ground practical reasons in some cases, there is something to the idea that certain desires (e.g., immoral or obsessive-compulsive desires) never justify an action. Perhaps Owen’s preference of a medical to a military career is sufficient to give him a reason to become a doctor. But what if Owen has a psychotic lapse, and is suddenly overcome by a desire to kidnap and murder small children? What if he develops an obsession-compulsive disorder, and finds himself gripped by an irresistible desire to count every star in the night sky? Does Owen really have reasons to give in to such strange and iniquitous desires? One wonders how the instrumentalist can say so without blushing. The intuition is that there are some ends for which it would be categorically irrational to desire or pursue. Let us call this intuition the practical fallibility of ends. Opponents of instrumentalism often tout the fallibility of ends
as decisive proof that instrumentalism is false. But in section 4.4 of this chapter, I suggest that anti-instrumentalists declare victory much too quickly.

4.2 External reasons and the practical fallibility of ends

There’s a palpable tension between the subjectivity of practical reasons and the practical fallibility of ends. On the one hand, different people’s idiosyncratic and subjective reactions to things seem to entail different practical reasons. In section 4.1, it was argued that at least some of the differences in reason-giving subjective reactions across agents stem from differences in desires. Instrumentalism can readily account for this fact. However, the instrumentalists are committed to a much stronger claim: namely that all inter-personal differences in practical reasons are attributable either to inter-personal differences in desires, or optimal ways of satisfying them. This raises the problem of the practical fallibility of ends. For the instrumentalist, if agent S maximizes the satisfaction of his desires by doing A in order to achieve end E, then S’s best reason favors doing A. But what if we substitute sadistic pleasure for “E,” and torturing toddlers for “A”? Critics of instrumentalism maintain that there must be normative limits to the subjectivity of reasons, since there are some ends which no person could rationally desire or pursue. But because instrumentalism cannot justify any such limits, it must be false. In the present section, I begin to argue that instrumentalism actually provides the most satisfactory reconciliation of the subjectivity of reasons and the fallibility of ends. Specifically, I claim that accounts of practical rationality which exclude some ends from being rationally pursuable fail to accommodate the subjectivity of reasons.
To fix ideas, it will help to return to Williams’s distinction between internal and external reasons for action (cf. chapter 1, section 1.4.3). *Internal reasons* are practical reasons which are derivable, through procedural practical reasoning, from the actual motivations of the agent. In this sense, they’re “internal” to an agent’s motivations. *External reasons* are practical reasons that agents have independently of their actual motivations.² Actual reasons internalists hold that all practical reasons are internal reasons. Instrumentalism is a variety of actual internalism, for instrumentalism holds that all practical reasons are instrumental reasons. And because instrumental reasons are constituted by the agent’s justified beliefs about how to satisfy his or her actual desires, all instrumental reasons are internal reasons. On the other hand, counterfactual reasons internalists and reasons externalists argue that at least some practical reasons are external reasons.³ Typically, philosophers of these stripes hold that external reasons bear a necessary relation to *objective values* concerning what is morally right, good, or intrinsically worth pursuing. They consider it a requirement of rational agency that these values be recognized as considerations that bear on what ought to be done. Thus, if an external reason happens to be an agent’s *best* reason for action, then either the agent acts

---

² In the literature, internal and external reasons have also been called “hypothetical” and “categorical” reasons, respectively (see Cullity & Gaut, 1997, pp. 3 – 4; Dreier, 1997; 2001). Internal reasons are “hypothetical” because they can be expressed as hypothetical (i.e., conditional) statements. The hypothetical component will be a condition that makes a statement about the agent’s actual motivations. Thus, internal reasons statements have the form “*S* has a reason *R* to *A* if *S* has some motivation to *A* because of *R*.” External reasons, on the other hand, are “categorical” because they are not contingent on agent’s actual motivations. Thus, external reasons statements have the form “*S* has a reason *R* to *A*.” Note here the absence of any “if” clause that makes reference to the agent’s actual motivations.

³ In chapter 1, I discuss at length the four-way distinction between instrumentalism, actual internalism, counterfactual internalism, and reasons externalism.
on this reason, or she is not practically rational (cf. chapter 1, section 1.5 – 1.6). Further, whether or not an external reason is an agent’s best reason for action does not depend on the agent’s actual motivations. Rather, external reasons are determined by substantive principles of practical rationality (according to counterfactual internalists—e.g., Korsgaard), or by the requirements of the situation (according to reasons externalists—e.g. McDowell). For present purposes, I shall overlook the subtle differences between counterfactual internalism and reasons externalism, and speak of both perspectives as theories of external reasons.\footnote{I also overlook McDowell’s efforts to undermine the distinction between internal and external reasons (see chapter 1, section 1.6). In my view, these efforts do not negate the most important consideration in favor of the distinction, as McDowell himself holds that reasons for action are not dependent for their existence on the motivations of any particular agent.}

Philosophers who stress the fallibility of ends argue that in some situations, certain ends can never be rationally desired or pursued. Those who hold this view effectively hold that there are external reasons against acting on certain ends. Theories of external reasons do not admit the possibility of an agent’s having a best reason to do pointless, self-destructive, or immoral deeds. At first blush, this is a distinct advantage of external reasons theories. Intuitively, it seems that nobody can have a reason to turn on every radio he sees, no matter how much he enjoys doing it. Such behavior is a waste of time, and there are better reasons to do more productive things. Likewise, it may seem that a sports fanatic cannot have a reason to commit suicide just because his favorite basketball team was eliminated from this year’s playoffs, no matter how convinced he is
that life has lost all meaning. A life is much too valuable a thing to waste on a basketball
game, and we have compelling (external) reasons against sacrificing what is important
for what is not. In addition, it may seem that a serial killer can never have a best reason to
kill an innocent person in cold blood, despite whatever sick satisfaction it brings him. To
many, it is beyond dispute that there are practical reasons which favor doing what is
morally right, and that people have reasons to act morally regardless of whether or not
they themselves care about being moral.

This last example underscores the most significant attraction of the idea of
external practical reasons. In chapter 3, section 3.1, I took note of the thesis of moral
rationalism. This is the claim that if an agent morally ought to do something, then he or
she has a reason (typically, a best reason) to do it. In the same section I presented the
thesis of moral absolutism: the idea that moral obligations apply to all people, regardless
of whether they’re actually motivated to act on those obligations. The notion of an
external reason for action is a direct consequence of these two claims. For if both moral
absolutism and moral rationalism are true, then everyone has reasons to act morally
regardless of their actual motivations. External reasons are just the sort of practical
reasons said to exist independently of people’s motivations. So, if there are any reasons to
act morally, they must be external reasons.

Although the notion of external reasons has some basis in intuition, in chapters 1
through 3 I’ve presented arguments to suggest that there are no external reasons. Chapter
2 in particular offered support for a central premise in Bernard Williams’s argument for
actual reasons internalism. This was Williams’s explanatory constraint—namely, the idea that if there are normative practical reasons, then it must be psychologically possible for them to motivate agents. But external reasons are by definition supposed to exist independently of whether or not they motivate agents. Thus, if Williams’s explanatory constraint holds true, there are no external reasons (cf. my reconstruction of Williams’s argument in section 1.4.3). Furthermore, in chapter 3, I argued that insofar as people are capable of acting morally, the best theory of what motivates moral conduct is a Humean account according to which moral action is motivated by instrumental reasons. Thus, if there are reasons to be moral that are capable of motivating agents, these reasons must be instrumental (hence internal) reasons.

On the other hand, a number of theorists maintain that there are external reasons for action. In support of their views, they emphasize the practical fallibility of ends. As we saw in this section, primary evidence for practical fallibility of ends comes from intuitions about thought experiments like the ones presented above: e.g., compulsively turning on radios, committing suicide over a basketball game, murdering innocent people, etc. In response, reasons internalists will try to explain these cases away by portraying them solely as violations of internal reasons for action. This, indeed, will be the rebuttal I begin to develop in the present chapter. But before presenting my positive argument, I will criticize an important theory of external reasons that I wish to distinguish from my own, instrumentalist view.
4.3 Practical rationality and human flourishing

4.3.1 Aristotelian roots

Philippa Foot was once an articulate proponent of actual reasons internalism. Then, in the 1990s, she repudiated her former writings on practical rationality and came to adopt a theory of external reasons (see Foot, 1994; 1995). Foot’s recent view is developed most systematically in her book *Natural Goodness* (2001). That work is essentially an update of Aristotle’s attempt, in the *Nicomachean Ethics*, to ground a theory of practical rationality on an account of human nature. I shall preface my discussion of Foot’s current view by emphasizing the parts of her thought that have an Aristotelian pedigree.

Contemporary philosophers who write about practical rationality tend to consider themselves heirs to one of three historical “greats”: Hume, Kant, or Aristotle. Instrumentalism is often characterized as a Humean theory of reasons, although some of Hume’s interpreters have argued that he was in fact a skeptic about practical reason (see especially Millgram, 1995). Christine Korsgaard and Thomas Nagel, the two most distinguished counterfactual internalists, pay tribute to Kant. But Foot and a number of other thinkers—including Martha Nussbaum and Rosalind Hursthouse—write in the shadow of Aristotle. Therefore, to establish the background for Foot’s new theory, it will be useful to briefly overview the “neo-Aristotelian” approach to practical reason (see especially Cullity & Gaut, 1997).
Most importantly, Aristotelians conceive of practically rational action as being conceptually dependent upon what is good for the agent. In other words, an agent has a reason to act in a certain way just in case it would be good for that agent to perform that action. Aristotelians focus on establishing a theory of an agent’s good, and from this they derive a notion of practical rationality. Following Aristotle himself, they treat practical rationality as a normative framework which applies specifically to the actions of human beings. Consequently, the relevant conception of the good on which the Aristotelian model of practical rationality is based is that of the human good. In Book I of the Nicomachean Ethics (hereafter abbreviated “NE”), Aristotle defends an account of the human good which he bases on considerations about the function, or ergon, of human beings (cf. NE, I.7; Cullity & Gaut, 1997, p. 13; Kraut, 2002, p. 64). This appeal to human function remains the cornerstone of the Aristotelian approach to practical reason.

The ergon of a thing is what it does that makes it what it is (Nagel, 1972, p. 253). A knife, for example, has the ergon of cutting. Cutting is what makes a knife a knife. In Aristotle’s view, a thing can be evaluated by how well it performs its ergon. Hence a good knife is one that cuts well. The virtues of a thing are those properties which make it perform its ergon well. Thus sharpness is a virtue of a knife, since this is a property which enables knives to cut well. It follows generally that whether or not an object or individual is a good thing of a certain type depends on its possession of virtues characteristic of that type. Aristotle applies these considerations to human beings as part of a “function argument” for a particular account of the human good. He contends that
the human *ergon* is “the soul’s activity that expresses reason or requires reason” (*NE*, I.7, 1098a7 – 8). Humans beings, in Aristotle’s view, are unique among living creatures insofar as they possess certain faculties of reason. It’s important to keep this claim in perspective, since Aristotle never denies that non-human animals possess *some* reasoning ability. According to Richard Kraut, Aristotle recognizes at various points in his oeuvre that animals exhibit capacities for thought, learning, and even practical intelligence (Kraut, 2002, p. 68). Aristotle’s precise view, then, is that there are specific faculties of reason which are possessed by human beings alone. Kraut lists them as follows: “…deliberating about what to do, looking for and demanding reasons for and against proposed courses of action, assessing our emotional responses and altering their strength, investigating why the world appears as it does, developing theories that explain the facts, arguing that certain theories give better explanations than others” (Ibid.).

Since engaging in reason is the human *ergon*, a good human is one who can reason well. For Aristotle, the human virtues are those dispositions of judgment and feeling which enable human beings to reason well and act on the basis of successful reasoning. Aristotle distinguishes between two types of human virtue: the *intellectual* virtues, which pertain to practical and theoretical reasoning, and the *ethical* virtues, such as justice, temperance, and courage (Kraut, 2007). The virtuous exercise of reason allows one to attain the highest good in human life. Aristotle identifies this highest human good by the term “eudaimonia,” which translates as “happiness,” “well-being,” or “flourishing” (*NE*, I.4, 1095a14 – 20; Kraut, 2002, pp. 52 – 53). *Eudaimonia*, Aristotle
tells us, is something that people generally take to be of supreme importance. It is an end
toward which all human activity is, and ought to be, directed at attaining. Furthermore,
*eudaimonia* is a final end rather than an instrumental end (cf. chapter 1, section 1.1). It is
an end desired for its own sake, and not for the sake of achieving some further end. In
fact, Aristotle suggests that *eudaimonia* is the *sole* final end in human life. It is that which
makes all good things good (*NE*, I.4, 1095a28). As a result, Aristotle often refers to
*eudaimonia* as the good, not merely a good among others.

Practical rationality, or, in Aristotle’s terminology, “practical wisdom” is reason
concerned with how to attain what is good for a human being—i.e., how to attain
*eudaimonia* (*NE*, VI.5, 1143b3 – 4; Cullity & Gaut, 1997, p. 14). The ideally practically
rational agent, which Aristotle calls the *phronimos*, is one who excels in his or her
powers to discern the good through practical reasoning and acts accordingly. *Phronimoi*
are able to reason well because they possess virtue, and by possessing virtue they have
reached a state of *eudaimonia*.

There has been some controversy about the patterns of thought which Aristotle
believes to constitute practical reasoning. Occasionally it sounds as though Aristotle
holds a purely instrumental view of practical reason. In one rather infamous passage, he
writes (*NE*, III.3, 1112b11 – 19):

> We deliberate not about ends, but about what promotes ends…we first lay down the end, and then examine the ways and means to achieve it. If it appears that any of several <possible> means will reach it, we consider which of them will reach it most easily and most finely.
However, Aristotle does not confine practical deliberation to means-ends reasoning. In particular, he describes another rational faculty which sets humans apart from the rest of the animal world. This is the capacity for one’s appetites and emotions to be governed by the demands of practical reason. Assuming that our appetites and emotions determine what ends we desire, as is likely, our ends can be evaluated and revised through practical deliberation in accordance with the requirements of reason (cf. Kraut, 2002, p. 69). That Aristotle held this view is most apparent when we consider the contribution of the ethical virtues to practical reasoning. According to Aristotle’s “doctrine of the mean,” every ethical virtue is a state intermediate between two extremes—one a deficiency, the other an excess (NE, II. 7, 1106a26 – b28). For example, bravery is a virtue intermediate between feelings of fear and overconfidence. A person who feels a deficiency of confidence is cowardly, and flees at every hint of danger. On the other hand, a person who feels an excess of confidence is reckless. It’s easy to see how a cowardly person might fail to flourish, since such a person would never take a risk for the sake of worthy cause. But a reckless person is also unlikely to flourish—he or she would be prone to taking extreme risks, and would probably end up paying for them. The virtue of bravery, however, is a disposition of thought and feeling reflected in an agent’s ability to recognize the “amount” of confidence which will most likely lead to success in a given situation (Kraut, 2007). But in order to exercise the virtue of bravery, one must be able to

---

suppress both fearful emotions and overconfidence. For Aristotle, then, practical reasoning encompasses the ability to judge the right balance of emotions one should have in order to flourish, as well as the ability to act on these judgments.

Aristotle’s doctrine of the mean suggests that achieving *eudaimonia* is not simply a matter of fulfilling one’s desires. The coward might have no inclination to rid himself of his tendency to flee from every sign of danger, and the reckless individual might thoroughly enjoy his excessive risk-taking. It’s conceivable that each of these characters acts in a manner consistent with his strongest desires. Nevertheless, Aristotle holds that *only* an intermediate state between cowardice and recklessness conduces to flourishing, and the ethical virtue of bravery enables a practically rational agent to reach this intermediate ideal. In sum, Aristotle’s account of *eudaimonia* explains the practical fallibility of ends. Regardless of how strongly an agent desires something, her desires may be detrimental to her flourishing. In that case, if she acts on these desires, she’ll fail to lead a good human life. All practically rational humans have reason to pursue their own flourishing, and yet flourishing is not contingent upon the desires of an agent. Therefore, Aristotle is committed to the existence of external reasons for action. Contemporary Aristotelians like Foot are with him on this.
4.3.2 Natural goodness

Foot’s recent theory of practical rationality shares the broad contours of Aristotle’s. Like Aristotle, Foot assumes that one’s reasons for action are determined by facts about what promotes the agent’s good. As Foot sees it, “there is no criterion for practical rationality which is not derived from that of goodness of the will” (Foot, 2001, p. 11). Additionally, for both Foot and Aristotle, the good of a human agent is constituted by facts about what is good for any human being generally. Foot follows Aristotle in endorsing a conception of the human good which is not dependent on the motivations of any particular individual. Furthermore, she conceives of virtues as dispositions which enable agents, insofar as they possess them, to do what is good—that is, to do what conduces to human flourishing (Foot, 2001, p. 12). She also tells us that practically rational agents are necessarily virtuous. Those who possess virtue are capable of judging certain considerations as reasons for acting, and then acting on these judgments. Included among the virtues are the ethical virtues, such as justice. The just person recognizes reasons to keep his own promises, to pay what is owed, to defend the rights of others, and so on (Ibid.). Upon recognizing these considerations, the just person will act on them. And insofar as the just person recognizes these considerations, he or she is practically rational.

---

6 In On Virtue Ethics (1999), Rosalind Hursthouse defends a position very similar to Foot’s. In many ways, Hursthouse’s theory has more structure and detail. Here I focus primarily on Foot, if only because her account is easier to summarize. Although there are nuanced differences between Foot and Hursthouse, I believe my criticisms of Foot also apply to Hursthouse’s view.
Foot also shares with Aristotle the view that there are external reasons for action. Both hold that an agent, $S$, has a practical reason to do $A$ if and only if $A$ is good for $S$. But if the human good is independent of the actual motivations of any particular agent, then one’s reasons for action are independent of the actual motivations of any agent. Again, Foot only recently came to this view. At the time she published her best-known essay, “Morality as a System of Hypothetical Imperatives,” she was wedded to an instrumental theory of reasons. In that piece, Foot argued that if there are any reasons to act morally, they must be instrumental reasons based on the actual desires and interests of the agent. Years later, in a stunning recantation, she abandons instrumentalism (see in particular Foot, 1994). By her own account, Foot acquiesced to arguments in favor of moral rationalism. She confesses to being persuaded by Warren Quinn’s contention that practical rationality is a master virtue (Foot, 2001, p. 62; Quinn, 1993). Quinn supposes that there are necessary conceptual connections between acting virtuously, acting rationally, and acting well. But if practical reason is concerned only with the relation between means and ends, it would rationalize all manner of vice—nastiness, brutality, etc.—for those who harbor vicious ends. Surely, Quinn contends, no person who acts immorally acts well. Therefore, no such person could act with full rationality, either. It follows that moral conduct is a requirement for being practically rational. But even since her younger days, Foot has been a moral absolutist. That is, she has believed all along that moral requirements apply to all persons, and do not vary with respect to any one person’s particular motivations. If moral requirements are absolute in this sense, and they
entail practical reasons, then Foot’s newfound moral rationalism commits her to the view that reasons to act morally are external reasons.

Foot intends for her theory to provide a normative framework which can justify claims about external reasons for action. Pursuant to this task, she deploys a line of thinking that echoes Aristotle’s “function argument” for the human good (cf. Kraut, 2002, pp. 63 – 70). Like Aristotle, Foot invokes the notion of function to support a conception of the objective human good. Here is Foot’s central thought: living things are subject to normative evaluations that are based on facts of biology. The evaluative predicates natural goodness and natural defect play a central role in Foot’s theory (see Foot, 2001, pp. 26 – 27; 30 – 34; 51). Natural goodness is the degree to which an organism’s parts, characteristics, and operations contribute to its life cycle. Natural defect, one the other hand, is the degree to which an organism’s parts, characteristics, etc. detract from its life cycle. The life cycle of an organism consists of the characteristic ways in which typical members of the biological species to which that organism belongs develop, survive, and reproduce within their natural environment (Foot, 2001, p. 33). For instance, the life cycle of a frog consists of hatching from an egg, developing from a tadpole into a full-grown adult, avoiding predators, catching food, mating, and (for females) laying eggs. The life cycles of other animals, like mammals and birds, include the rearing of young as part of the process of reproduction.

Foot clarifies her concepts of natural goodness and defect by applying them to the relatively simple cases of plants and non-human animals. Consider a peacock without its
colorful plumage. It seems right to say not only that this peacock is unusual, since most peacocks are endowed with beautiful tail feathers, but also that he has a defect. What licenses us in calling this peacock defective is that his lack of brilliant tail feathers undermines his ability to attract a mate and reproduce. By contrast, a peacock with a brilliant tail can naturally be described as a good peacock. The leading explanation for this seems to be that a peacock with an exceptionally colorful tail will be more successful at reproducing.

Foot also points out that natural goodness and defect do not refer to the statistically normal traits of a species (Foot, 2001, pp. 30 – 31). That a peacock without a tail is naturally defective is not explained by the fact that he is different from most peacocks. For suppose one day during the mating season the tail feathers of all living peacocks fell off inexplicably. Would a peacock without a tail then not be defective? No. Instead, it seems that any tail-less peacock is naturally defective, since being tail-less undermines a peacock’s chances for reproduction.

Foot’s concepts of natural goodness and defect are derived from assumptions about the functions served by the various features of an organism (Foot, 2001, pp. 32 – 33). She suggests that the function of a trait is what it contributes to the life cycle of the trait-bearer (Foot, 2001, p. 32). Let’s call anything which falls under this definition an “F-function” (with the “F” standing for “Foot”). The F-functions of a creature’s traits

---

7 The concept of function has played a fascinating myriad of roles in the history of science. For an overview, see Woodfield’s (1998) article on teleology in the Routledge Encyclopedia of Philosophy.
establish what Foot calls a “schema of natural normativity”—i.e., a set of norms on the basis of which judgments about the goodness or defectiveness of a living thing are made (Foot, 2001, pp. 33 – 34, 40). For example, the F-function of a peacock’s tail is to attract mates, and hence increase the peacock’s chances for reproduction. If his tail is especially apt for its F-function—for instance, if it’s particularly colorful—then the tail is a natural good. If the peacock’s tail serves its function poorly—e.g., if it were drably colored—then the tail is a natural defect.

Interestingly, Foot insists that her concept of an F-function is not the concept of function as it’s understood in evolutionary biology (see especially Foot, 2001, p. 32, n. 10). The orthodox account of function among evolutionary biologists is known as the etiological theory. According to this theory, the biological function of a trait is the effect of that trait which explains why its bearers were favored by natural selection in the past (Sterelny & Griffiths, 1999, p. 223; Millikan, 1989, p. 289). Etiological theorists—in particular, Ruth Millikan—have called functions which meet this description proper functions. Foot never explains in detail how her F-functions differ from proper functions, though she suggests that her concept is more compatible with “everyday uses” uses of the term “function” (Foot, 2001, p. 32, n. 10). Judging from her comments, however, Foot seems to want strictly to define F-functions in terms of what a trait contributes to a creature’s life cycle, and not in terms of natural selection.

---

8 For an introduction to the basic perspectives on biological functions, see Cummins, 1975; Millikan, 1989; Neander, 1991; Griffiths, 1993; Sterelny & Griffiths, 1999 (§10.2).
9 I already cited the idea of proper function in chapter 2, section 2.3.
It’s important to see that F-functions apply to types of living thing, rather than tokens. The F-function of a trait is not something which promotes the development, survival, and reproduction of a particular trait-bearer. Instead, F-functions belong to the traits of typical members of a species of organism living in their natural habitat. Foot offers the example of a slow-footed deer living in a zoo (Foot, 2001, p. 34). Since the deer is not hunted in this man-made habitat, its slowness doesn’t diminish its ability to survive. Nevertheless, according to Foot this animal is still naturally defective. This is because being slow of foot reduces chances for survival for a typical member of that deer’s species living in its natural habitat, where predators abound.

Foot’s definitive move is to draw an analogy between the criteria for natural goodness and the criteria for moral goodness. In particular, she seeks to ground an account of the moral virtues in considerations about the human life cycle. Foot could hardly be more up-front about her agenda: “I am,” she proclaims, “quite seriously, likening the basis of moral evaluation to that of the evaluation of behaviour in animals” (Foot, 2001, p. 16). Consider these remarks, where Foot lays down her cards in more detail (Foot, 2001, p. 24):

In my view…a moral evaluation does not stand over against the statement of a matter of fact, but rather has to do with facts about a particular subject matter, as do evaluations of such things as sight and hearing in animals, and other aspects of their behaviour. Nobody would, I think, take it as other than a plain matter of fact that there is something wrong with the hearing of a gull that cannot distinguish the cry of its own chick, as with the sight of an owl that cannot see in the dark. Similarly, it is obvious that there are objective, factual evaluations of such things as human sight, hearing, memory, and concentration, based on the life-form of our own species. Why, then, does it seem so monstrous a suggestion that the
evaluation of the human will should be determined by facts about the nature of human beings and the life of our own species?

In this passage, Foot commits herself to a form of moral naturalism (Lenman, 2006). On any such view, there are objective moral facts, and these are facts about the natural world. Foot’s moral naturalism rests on an argument by analogy. Her primary datum is that evaluations of natural goodness and defect seem capable of being objectively true. An owl that cannot see in the dark is a defective owl, and Foot takes this judgment to express an obvious and objective fact. The truth-maker for evaluations of natural goodness and defect is the extent to which a living thing’s parts and operations are performing their F-functions successfully. If so, then they promote the life cycle of the organism. And to gain knowledge about the life cycles of organisms, we need look no further than biology.

So, the first premise in Foot’s argument by analogy is that some evaluations of natural goodness are objectively true. The next stage in Foot’s argument is to identify a relevant respect in which evaluations of natural goodness are similar to moral evaluations of persons. She emphasizes what she calls “other-regarding goodesses” in non-humans (Foot, 2001, p. 35). These are behavioral dispositions to help or cooperate with others. Examples offered by Foot include the dance of the honeybee, which informs other bees of a source of food, the sting of the bee, and pack-hunting in wolves. Other-regarding goodesses are natural goods because they are part of the species-specific life cycles of the animals which engage in them (Foot, 2001, p. 16). Individual bees and wolves that engage in these behaviors are naturally good, because such behavior conduces to survival and reproduction. A sterile worker bee that defends the hive with its suicidal sting helps
the queen, with whom the worker shares roughly three quarters of its genes, to continue producing offspring. Therefore, a bee that stings is a naturally good bee. On the other hand, a bee that does not defend the hive is naturally defective, because she does not reproduce her genes as effectively as she could by using her sting. Further examples of natural defect due to failure to engage in other-regarding goodesses include a free-riding wolf that feeds on the group’s kill without contributing to the hunt, or a bee that finds a source of nectar but does not signal its location to others.

Following Peter Geach, Foot compares other-regarding goodesses in animals to moral virtues in humans. She says at one point that the “virtues play a necessary part in the life of human beings as do stings in the life of bees” (Foot, 2001, p. 35). Later, Foot describes the parallel in fuller detail (Foot, 2001, pp. 44 – 45):

…the concept of a good human life plays the same part in determining goodness of human characteristics and operations that the concept of flourishing plays in the determination of goodness in plants and animals…Nor is there any reason to think that it would not be in place even in the evaluations that are nowadays spoken of as the special domain of morality…Men and women need to be industrious and tenacious of purpose not only so as to be able to house, clothe, and feed themselves, but also to pursue human ends having to do with love and friendship. They need the ability to form family ties, friendships, and special relations with their neighbors. They also need codes of conduct. And how could they have all these things without virtues such as loyalty, fairness, kindness, and in certain circumstances obedience?

According to Foot, a plant or nonhuman animal flourishes by engaging in those activities that contribute to its life cycle—i.e., to its development, survival, and reproduction. In animals, cooperative relations with others contribute to the individual’s survival and reproduction. An animal that engages in these cooperative relations is naturally good. But
clearly, cooperative relations with others are also a central part of the human life cycle. Foot presses this point with the example of promise-keeping (Foot, 2001, pp. 45 – 51). She quotes from Elizabeth Anscombe’s paper, “On Promising and Its Justice” (Anscombe, 1981). There, Anscombe observes that people rely on others to attain the things they need. We cannot secure all our needs alone, just as a lone wolf would not be as successful at capturing prey as a wolf that hunts with the pack. However, Anscombe also points out that few human beings have the power to coerce others to do the things they need by force. So instead, we’ve developed the institution of promising to recruit one another’s assistance. We can attract the help of other people by making credible promises to help or benefit the other in kind, at some point in the future. But to make credible promises to reciprocate, one has to be trustworthy. And to win the trust of others, it’s important that one be respectful (cf. Foot, 2001, p. 48). Therefore, respectfulness and trustworthiness are two dispositions that enable human beings to secure the benefits of cooperation. And insofar as cooperative interactions are a necessary part of the human life cycle, respect and trustworthiness contribute to that life cycle.

It follows, on Foot’s account, that a human being who is respectful and trustworthy is naturally good. Furthermore, respect and trustworthiness meet the traditional Aristotelian definition of a virtue; they are traits of character which promote the flourishing of the agent. All told, cooperative relations with others contribute to the life cycles of humans and some non-human species. Non-human animals have traits that enable them to engage in cooperation—e.g., the bee’s sting or the wolf’s pack instinct. A
non-human that bears these traits will therefore be a (naturally) good specimen of its kind. Human beings, too, have traits that promote cooperative relations with others. These traits are the moral virtues, and they include respect, trustworthiness, kindness, loyalty, and the like. A virtuous human being is therefore a naturally good one—by possessing the virtues, he or she can flourish. Thus, evaluations of natural goodness in nonhumans are similar to moral evaluations of humans in this sense: both modes of evaluation are justified by the standard of contributing to the characteristic life cycle of a species.

Having identified a relevant similarity between natural goodness and moral goodness, Foot concludes that moral evaluations of human beings with respect to their possession of virtue are just as objective and factual as evaluations of natural goodness in nonhumans. From this conclusion, Foot draws consequences for a theory of practical reason. Like Warren Quinn, she suggests that the very concept of possessing a virtue involves acting and choosing well (Foot, 2001, p. 12). Thus an account of virtue is an account of good action, and an account of good action is all that we would expect from a theory of practical rationality. On this matter, Foot is true to her Aristotelian roots: any human being has reason to do what it takes to flourish, and flourishing is achieved by action in conformity with virtue.

4.3.3 Watson’s dilemma

In my view, the theory of practical reason put forth by Foot and like-minded Aristotelians falls to a resilient objection voiced by Gary Watson in his paper, “On the
Primacy of Character” (in Statman, 1997). Watson raises a challenge against the variant of neo-Aristotelian virtue ethics championed by the likes of Foot. As we’ve seen, Foot’s project is to base morality on some objective and naturalistically grounded theory of human nature. Watson states his doubts about this enterprise in the form of a dilemma (Watson, in Statman, 1997, p. 67):

Either the [virtue ethicist’s] pivotal account of human nature (or characteristic human life) will be morally indeterminate, or it will not be objectively well founded…An objective account of human nature would imply, perhaps, that a good human life must be social in character. This implication will disqualify the sociopath but not the Hell's Angel. The contrast is revealing, for we tend to regard the sociopath not as evil but as beyond the pale of morality. On the other hand, if we enrich our conception of sociality to exclude Hell’s Angels, the worry is that this conception will no longer ground moral judgment but rather express it.

Watson construes virtue ethics as a family of ethical theories influenced primarily by Aristotle. But his critique is equally threatening to the Aristotelian view of practical rationality. For the Aristotelians, the requirements of ethics and the requirements of practical reason come to the same thing—both are requirements of a life lived in conformity with virtue. The thrust of Watson’s doubt is this: virtue ethicists purport to give an objective theory of the human good based on considerations about human nature. Watson, however, thinks that the virtue ethicist’s characterization of human nature fails to provide a determinate and intuitively satisfying specification of the virtues. On the other hand, when the neo-Aristotelian tries to provide such a specification, it comes unmoored from its foundations in putatively objective truths about human nature. In sum, Watson’s dilemma can be expressed as the following charge: virtue ethics is either indeterminate or it lacks objective foundations. The root of these complications, I shall
ultimately suggest, lies in subjectivity of reasons. Neo-Aristotelians have failed to develop an adequately nuanced characterization of human nature which is compatible with the intuition that reasons for action can vary with the individual.

Watson’s dilemma can be illustrated with an example provided by Foot herself, as part of an argument for the goodness of keeping promises (Foot, 2001, pp. 47 – 50). She contends, through her discussion of Anscombe, that promise-keeping is good because it promotes cooperative interactions which make up a central component of the human life cycle. However, Foot also acknowledges that there will always be particular situations in which keeping or breaking a promise has little or no impact on cooperation. She recounts Peter Kropotkin’s story about Mikluko-Maklay, the Russian geographer who studied the indigenous inhabitants of the Malayan archipelago. Maklay employed a native who agreed to provide his services on the condition that Maklay would never photograph him. But one day, while his employee was fast asleep, Maklay was beset by the temptation to break his promise. From Maklay’s point of view, his native friend would have been an ideal subject. Moreover, he never would have found out about being photographed. So why not take the photo?

Maklay ultimately honored his agreement. The question is whether it would have been wrong—i.e., contrary to virtue—for Maklay to take a picture of his aid. Intuitively, it does seem as though Maklay would’ve acted wrongly by breaking his promise. But is this intuition supported by Foot’s theory of natural goodness? For the following reasons, this is doubtful. Foot wants to distance her own view from some type of consequentialism
which requires the agent to do whatever will produce good results (Foot, 2001, pp. 47 –
49). In the context of Foot’s theory and the Maklay example, “good results” could be
interpreted as the development, survival, or reproduction of a typical human being in his
or her natural environment. If Foot’s theory were consequentialist in this sense, then the
goodness or badness of Maklay’s breaking his promise would depend on whether or not
keeping the promise would conduce to a typical human being’s development, survival,
and reproduction. However, it’s not clear that Maklay’s choice of breaking or keeping
this particular promise would significantly affect his ability to secure these ends. On the
one hand, perhaps the native would have been so offended that he would quit working for
Maklay. In that case, Maklay would be risking the benefits that flow from the institution
of promising—benefits which may carry meaningful implications in terms of survival or
reproduction. But by hypothesis of the example, Foot acknowledges that any such risk is
negligible in Maklay’s situation. On the other hand, Foot even acknowledges that on a
consequentialist interpretation of her theory, it may have turned out to be good for
Maklay to break his promise. Perhaps there are moral virtues associated with scholarly
work—love of truth, for instance—which contribute to the human life cycle (Foot, 2001,
p. 50). If a photograph of his native aid would contribute to the anthropological record,
then wouldn’t Maklay be acting virtuously by taking the picture? It’s exceedingly
difficult to spell out in any precise way the relevant risks and tradeoffs associated with
Maklay’s choice in terms of its potential effects on the human life cycle. Indeed, whether
Maklay chose to keep his promise or not, it does not seem his choice would have a
significant enough impact on development, survival, or reproduction to support a determinate moral evaluation.

Intuition suggests that when Maklay chose to keep his promise, he chose rightly. And in her assessment of Maklay’s case, Foot sides with intuition. She tells us that Maklay would have acted wrongly by taking the photo. But in justifying this intuition, Foot appeals to considerations which seem entirely foreign to her theory of natural goodness. Her main point is the following (Foot, 2001, p. 51):

In giving a promise one makes use of a special kind of tool invented by humans for the better conduct of their lives, creating an obligation that (although not absolute) contains in its nature an obligation that harmlessness does not annul.

The thought, here, is that the institution of promising contributes to human flourishing. Trustworthiness and respectfulness are virtues that people need in order to for this institution to be maintained. The act of making a promise creates obligations that are not expunged whenever it turns out that breaking a promise will bring about good consequences. Whatever good things might have come from photographing his servant, Maklay would have exhibited disrespect and untrustworthiness by breaking his promise, and according to Foot, “disrespect and untrustworthiness are bad human dispositions” (Foot, 2001, p. 48). In this way, Foot is committed to a non-consequentialist ethics. Human actions are not to be judged directly by how they contribute to the human life cycle, but rather by rules of conduct, institutions, and practices which conduce to human flourishing. We are morally good insofar as we have the disposition—i.e., virtues—
which enable us to safeguard and live in harmony with those rules, institutions, and practices.

However, it’s not at all clear how anything but a form of life cycle consequentialism can be supported by Foot’s account of natural goodness. The problem is that Foot’s basic examples of natural goodness and defect in plants and nonhuman animals strongly suggest in favor of a consequentialist ethic. Such a principle takes the optimal promotion of a species-specific life cycle as the criterion of goodness. A swift deer is a good deer because it’s more likely to survive in its natural habitat, whereas a slow deer is defective because predators can take it down with greater ease. It seems to follow straightforwardly from Foot’s account that if swiftness is part of that bundle of characteristics which promotes a typical deer’s ability to develop, survive, and reproduce in its natural environment, then it is good for a deer to be swift. But this same consequentialist principle might not support the judgment that Maklay would act wrongly by breaking his promise. For how would the breaking of this particular promise be in any way detrimental to the human life cycle? One possibility we already noted is that it would spoil future opportunities for cooperative interactions. However, even Foot concedes that Maklay didn’t have to worry about that, since his native assistant never would have known that he’d been photographed. Another possibility is that breaking a promise in this one isolated instance would somehow weaken Maklay’s general disposition to be trustworthy and respectful in future interactions where those virtues might prove beneficial. But wouldn’t Maklay do just as well, if not better, by cultivating a disposition
to keep-apromise-except-when-the-consequences-of-breaking-it-are-sufficiently-favorable? True, there is always some potential for error in identifying situations where promise-breaking is sufficiently favorable. But with respect to Maklay’s situation, there seems little room for doubt that the geographer would benefit by forgetting his promise and, in the name of science, taking one little photo of his native companion.

When all is said and done, Foot can bring her ethics into line with intuition only by abandoning its foundation in a theory of natural goodness. Her view thus collides with Watson’s dilemma. On the one hand, if Foot’s theory of natural goodness establishes an objective foundation for any ethics, it would have to be a form of life-cycle consequentialism. But this consequentialist ethic does not seem to support an intuitively plausible specification of virtuous action. On the other hand, when trying to defend an intuitive specification of the virtues, Foot fails to demonstrate how it can be derived from the supposedly objective account of human nature she offers.

The foregoing discussion of the Maklay example put weight on the credentials of Foot’s view as an ethical theory. But when we think of Foot’s view more directly as an account of practical rationality, Watson’s dilemma applies equally well. In some contexts, people’s choices have little or no moral significance. Even in such non-moral situations, practical reasons may favor some actions over others. Consider, for example, the choice of buying a car. I might be considering two equally expensive models: one has plenty of accoutrements and high customer-satisfaction ratings. Another gets great gas mileage and has an engine that will last. How could considerations relating to
development, survival, and reproduction favor my buying one model over another? Of course, some cars may be better designed to protect drivers in the event of a life-threatening collision. And some cars may be more suitable than others in attracting potential mates. But let’s suppose the two cars I am considering are equally safe and likely to impress the opposite sex. In fact, assume that it won’t make any difference to my prospects for development, survival, and reproduction if I buy one car or the other. In that case, reflecting on the human life cycle would seem to provide no further considerations to favor either car. However, it seems that there are considerations which would give me better reason to buy car A than car B. Among the most salient that come to mind are my preferences over mutually exclusive alternatives. Given what I can afford, perhaps I would be willing to give up a few luxuries for the sake of being able to drive long distances at a lower expense. This seems to provide sufficient reason for me to buy the more gas-efficient car, even if the choice of either car would not make a difference to my prospects for survival or reproduction. Once again, in line with the first horn of Watson’s dilemma, Foot’s concept of natural goodness does not provide a basis for making determinate and intuitively plausible judgments about one’s reasons for action, even in cases where it seems such judgments should be possible.

Now Foot never denies that at least in some cases, a person’s desires can play a role in determining the practical reasons he or she has. It may be that natural goodness only imposes broad constraints on the ends one may rationally pursue. Within the limits of those constraints, desire might play a role in determining one’s practical reasons. Yet
even when it comes to evaluating actions that bear directly on the ends associated with
natural goodness—development, survival, and reproduction—Foot’s theory of the human
good yields intuitively faulty results. According to Foot, the lack of capacity to reproduce
is a natural defect in a human being, since reproduction is part of the human life cycle
(Foot, 2001, p. 42). But suppose someone chooses to be childless. More to the point,
imagine someone who chooses to have no children despite being capable of bearing and
rearing quite a few of them. Would the choice not to meet one’s full reproductive
potential, then, be a defective and hence practically irrational one? Similar questions can
be asked about the choice of celibacy, or of homosexual behavior. Intuition, it seems,
favors the idea that people can have excellent reasons to bear fewer offspring than they’re
able to support. There are many worthy pursuits other than childrearing, such as a career
or spiritual calling, to which one could rationally devote oneself. For her part, Foot wants
to concur with this intuition, but it’s by no means obvious how it could be supported by
her theoretical account of natural goodness. Again, Foot’s ethics falls upon the first horn
of Watson’s dilemma.

To show how a human being’s choice of childlessness can be rational within her
framework, Foot seems compelled to propose a more expansive account of the human life
cycle which is not limited to developing, surviving, and reproducing. This move is
necessary because, on Foot’s view, natural goodness evaluations are to be made with
reference to the life cycle of a species. Observe the following comments (Foot, 2001, p.
42, my emphases):
Whether an individual plant or animal actually succeeds in living the life that it is its good to live depends on...the interaction of natural habitat and natural (species-general) ‘strategies’ for survival and reproduction. What conceptually determines goodness in a feature or operation is the relation, for the species, of that feature or operation to survival and reproduction...But clearly this is not true when we come to human beings...choice of childlessness and even celibacy is not thereby shown to be a defective choice, because human good is not the same as plant or animal good. The bearing and rearing of children is not an ultimate good in human life, because other elements of good such as the demands of work to be done may give a man or woman reason to renounce family life.

The passage suggests that the human life cycle, unlike animal life cycles, includes elements besides developing, surviving, and reproducing. She cites “work to be done” as one specifically human activity which may rationalize the choice of childlessness. This tack of introducing a more inclusive account of the human life cycle seems to be a promising way for Foot to avoid the first horn of Watson’s dilemma—that is, it seems to provide some basis for a plausible and determinate specification of rational action.

However, the price of this strategy is that it runs up against the second horn of Watson’s dilemma. It seems to me that an expansive account of the human life cycle, which includes elements like “work to be done,” fails to provide an objective foundation for natural goodness evaluations. Recall that Foot assessed natural goodness in plants and non-human animals with reference to how well their features and operations promote development, survival, and reproduction. These evaluations had the sound of objective truths. But as far as I can tell, a human life cycle conceived more broadly to include elements besides developing, surviving, and reproducing cannot provide objective support for natural goodness evaluations in human beings. This can be seen if we consider why Foot’s basic examples of natural goodness evaluations in bees, wolves, and
other nonhumans have their seemingly objective, factual status. Why, for instance, does it ring so straightforwardly true to say that a peacock without colorful tail-feathers is a defective peacock? Foot never answers this question, I think to her detriment. For in my view, natural goodness evaluations in nonhumans have their seemingly objective status because they are grounded in a limited conception of an organism's life cycle which only comprises activities related to development, survival, and reproduction.

The apparent truth of natural goodness evaluations—particularly when applied to non-human species—can be explained if we interpret them as statements about what Peter Railton has called naturalized norms (Railton, 1986, p. 185). Naturalized norms are normative criteria for how things ought to be that play a role in explaining natural phenomena. As an example, Railton asks us to think about a potential explanation for the collapse of a roof. The explanation may be expressed by the sentence “The rafters ought to have been 2 x 8’s at least, but they were only 2 x 6’s” (Ibid.). This explanation cites as the cause of the house’s collapse the failure of the structure to be as it ought to be. And yet it seems to be a perfectly acceptable explanation which may very well be true. What could make it true?

Railton thinks that claims about naturalized norms can be true because naturalized norms are real features of the world. Their reality is confirmed by the fact that they can be postulated to explain observable phenomena. For Railton, naturalized norms appear in criterial explanations. These are explanations which explain why something happened on the basis of how well some naturalized norm is satisfied. Return to the explanation of the
collapsed roof. The naturalized norm which figures in this explanation can be described by the following sentence:

\[(R) \quad \text{If a roof of that design is to withstand the snow load this one bore, then it ought to be framed with rafters at least 2 x 8 in cross-section.}\]

Criterial explanations posit a process that selects for things which satisfy a naturalized norm (Railton, 1986, p. 185). For example, the laws of physics make it the case that roofs built according to the naturalized norm described in (R) will be sturdy enough to bear a large weight of snow. On the other hand, roofs built with narrower rafters than the dimensions cited in (R) will tend to collapse. The satisfaction of a naturalized norm has observable effects. (R) identifies a criterion—viz., that of having rafters of certain dimensions—which is causally responsible for observable phenomena—viz., sturdy roofs. Departures from a naturalized norm have observable consequences, as well. Rafters with narrower dimensions than 2 x 8 depart from the norm specified in (R), and the consequence of this departure would be a collapsed roof. According to Railton, the fact that the criterion described by (R) can bring about these effects confirms that it is a really existing natural property. And because (R) identifies this criterion correctly, it is true. In sum, if a skeptic were to ask how it could be true that a roof *ought* to have rafters at least 2 x 8 in cross-section, Railton’s answer is that a roof with those dimensions would hold up.
Natural goodness evaluations can also be interpreted as claims about naturalized norms. These norms specify how an organism *ought* to be in order to develop, survive, and reproduce in its natural environment. Consider the following sentence, which describes a naturalized norm about web-spinning ability in orb-weaver spiders:

(F) If orb-weavers are to survive and reproduce in their natural environments, then they ought to spin webs that can effectively capture prey.

The naturalized norm described by this sentence can figure in a criterial explanation for the continued existence of orb-weavers. Orb-weavers are able to survive and propagate themselves, in part, because they can capture food with their webs. Orb-weavers that didn’t spin webs would not be able to survive and reproduce as successfully. This criterial explanation for the persistence of orb-weavers cites natural selection as the process which selects for web-spinning orb-weavers. Following Railton, I suggest that the naturalized norm cited in (F) is what underwrites the truth of evaluative sentences like “A good orb-weaver *ought* to spin webs.” Again, if a skeptic were to insist on an account of what makes this sentence true, the Railton-style response would be that orb-weavers that didn’t spin webs would die out while web-spinning orb weavers would thrive.

Foot would almost certainly reject this account of the truth of natural goodness evaluations. At first blush, it might appear that this account follows straightforwardly from Foot’s proposal that the natural goodness of an organism be assessed on the basis of
the whether or not its traits conduce to the life cycle of a species. However, we have seen that Foot thinks the life cycle of human beings is more complex than that of non-human creatures. When discussing her basic examples of natural goodness in non-humans, Foot limits the life cycle to the ways in which an organism develops, survives, and reproduces (see especially Foot, 2001, pp. 32 – 33). But later, Foot suggests that evaluations of human goodness should not be based on the ends of development, survival, and reproduction alone:

...human good must indeed be recognized as different from good in the world of plants or animals, where good consisted in success in the cycle of development, self-maintenance, and reproduction. Human good is *sui generis*. Nevertheless, I maintain that a common conceptual structure remains. For there is a ‘natural-history story’ about how human beings achieve this good as there is about how plants and animals achieve theirs. There are truths such as ‘Humans make clothes and build houses’ that are to be compared with ‘Birds grow feathers and build nests’; but also propositions such as ‘Humans establish rules of conduct and recognize rights’. To determine what is goodness and what defect of character, disposition, and choice, we must consider...what kind of a living thing a human being is (Foot, 2001, p. 51).

Judging from this passage and her comment that voluntary childlessness is not a defective choice, Foot suggests that a human being can fail to survive or reproduce optimally, but still not be naturally defective as long as he or she is successful in meeting other ends (e.g., the “demands of work”) characteristic of the human life cycle.

But this suggestion is incoherent. Foot’s whole argument for the objective status of moral evaluations rested on an analogy with the apparent objectivity of natural goodness evaluations in animals and plants. And yet, as the discussion of Railton’s naturalized norms revealed, it seems that the objectivity of these evaluations is ultimately
based on the extent to which an organism is equipped with traits which enable it to
develop, survive, and reproduce. Therefore, the claim that plants and animals \textit{ought} to
have certain characteristics can be justified by a counterfactual test: if organisms of type
$O$ did not have such-and-such characteristics, then $O$ would cease to exist. Orb-weaver
spiders ought to be able to spin webs because, if they didn’t have this ability, they would
die out. But this counterfactual test implies that the choice of childlessness \textit{couldn’t} be a
naturally good choice. If certain people did not choose to bear as many children as they
were capable of rearing to reproductive age, then their genes would be less likely to be
transmitted to future generations than the genes of people who have as many children as
they can. Therefore, meeting one’s full reproductive potential is a natural good, and the
choice of childlessness, at least in the case of someone who is able to rear a child
successfully, must be a natural defect. When Foot appeals to “the demands of work” to
rationalize the choice of childlessness, she seems to be smuggling in an entirely different
normative standard. Whatever that standard is, it is not the same one which underwrites
the objectivity of natural goodness evaluations in non-humans. To let this new standard
play a part in determining the goodness of human life has the advantage of making it
possible to see the choice of childlessness as a good or rational choice, and this keeps
with intuition. However, the new standard lacks a foundation in an analogy with
objective, factual status of natural goodness evaluations in plants and animals.

In sum, Foot’s recent theory falls afoul of Watson’s dilemma. On the one hand, it
fails to support determinate, intuitively plausible evaluations of conduct. On the other
hand, when Foot tries arguing in favor of the rationality of some specific choice, her conclusions are not clearly grounded in her account of natural goodness. Ultimately, the subjectivity of practical reasons causes Foot a great deal of trouble. Whether or not I have reason to purchase one of two equally safe, affordable, and attractive car models seems inevitably to come down to my own idiosyncratic tastes. Whether or not one has reason to take on the burdens and joys of parenthood is, to some degree, a matter of the specific priorities of the individual. Foot’s theory of natural goodness is both too vague and too procrustean to make sense of these powerful intuitions.

4.4 The instrumental rationality of ends

4.4.1 Balancing two intuitions: the subjectivity of reasons vs. the fallibility of ends

In section 4.3, I identified Foot’s approach to practical rationality as one which claims to establish a basis for evaluating the desires of agents. I rejected this theory on the grounds that it fails to take into account the subjectivity of reasons. I think this same criticism applies to other theories of external reasons, including Scanlon’s. Scanlon’s account does have the advantage of being able to explain the subjectivity of reasons in terms of people’s “subjective reactions” to things. But in section 4.1, I argued that subjective reactions could easily be interpreted along the lines of desires and preferences. Also, Scanlon’s claim that desires do not provide reasons for action conflicts with the intuition I drew from the case in which Owen Wingrave has to choose between two equally promising careers: medicine or the military. In that thought experiment, it
appeared that Owen had a better reason to pursue a career in medicine solely by virtue of the fact that he preferred it to a military career. Thus, by denying that desires provide reasons for action, Scanlon ignores an important facet of the subjectivity of reasons: reasons vary from person to person, to some extent, because people want different things. And if so, then instrumentalism can provide a very natural explanation for the subjectivity of reasons. On the other hand, we also noted the charge that instrumentalism fails to make sense of the practical fallibility of ends—i.e., the intuition that certain desires and ends can never be pursued rationally. In this section, I argue that instrumentalism is capable of striking an optimal balance between the subjectivity of reasons and the practical fallibility of ends. Contrary to the conventional wisdom, instrumentalism can provide a normative perspective from which to criticize the desired ends of agents. Even for instrumentalists, then, not all desires are created equal.

It’s easy enough to see how instrumentalism accounts for the subjectivity of practical reasons. According to instrumentalism, all reasons for action are instrumental reasons. Instrumental reasons are determined by the desires of the individual and the optimal means of achieving those desires. But obviously, different people can have different desires. And yet it’s exactly on this point that instrumentalism faces the most dogged resistance. Proponents of external reasons argue that there must be constraints on the ends rational agents can desire. In section 4.2, I called this thesis the practical fallibility of ends. Again, we can easily think of a serial killer who wants to brutally murder innocent people, or of a person suffering from obsessive-compulsive disorder
who wants to wash his hands a thousand times a day. External reasons theorists maintain that no rational agent would have or act on such desires. Here, intuition seems to be with the external reasons theorist. But under instrumentalism, it’s at least logically possible for an agent with deeply immoral, self-destructive, or obsessive desires to act rationally on them, so long as she maximizes the overall satisfaction of her desires in the process. External reasons theorists thus declare themselves victorious on the grounds that instrumentalism fails to appreciate the practical fallibility of ends.

These dismissals of instrumentalism have been too hasty. More sophisticated versions of instrumentalism are compatible with the intuition that we can evaluate and revise our desires. Instrumentalism is a strictly “means-ends” conception of practical rationality—the only normative standard it upholds for assessing the rationality of an action is whether the action is an optimal means to achieving one’s most strongly desired ends. However, as David Schmidtz argues in “Choosing Ends” (Schmidtz, 2001; see also Schmidtz, 1995), even a means-ends conception of practical rationality is capable of subjecting one’s desired ends to normative scrutiny. I have great admiration for Schmidtz’s argument, and I shall use it as a starting point to develop, in chapter 5, an even more detailed account of how ends can be assessed within a purely instrumentalist framework (see chapter 5, section 5.6.2). The key move is to supplement a means-ends conception of rationality with a descriptive psychological theory of what human beings
generally desire.\textsuperscript{10} Such a theory would support the judgment that certain kinds of actions are instrumentally irrational for most people to undertake, so far as they tend to thwart the achievement of ends that most human beings strongly desire. If, according to this theory, most human beings strongly desire \( E \), and some person acts in a manner that makes the fulfillment of \( E \) less likely, then we have strong evidence that the person in question fails to be practically rational. In this way, instrumentalism is consistent with the practical fallibility of ends.

This more sophisticated version of instrumentalism incorporates a key insight from the neo-Aristotelian approach to practical rationality. The suggestion, pressed by both Aristotle and Foot, was that practically rational action is determined in significant ways by facts about human nature. Instrumentalism is therefore compatible with the thought that the pursuit of ends which run contrary to the requirements of human life is open to rational criticism. At the same time, the sophisticated form of instrumentalism defended here retains a purely means-ends conception of practical rationality. Consequently, the practical reasons of an agent are still constituted by the desires of that agent and the best means for achieving them. Though true generalizations about the things human beings desire can be made, it remains the case that desires can vary across individuals in terms of their specific objects and their strengths. Therefore, the enhanced

\textsuperscript{10} In chapter 5, I shall call these general human desires the \textit{basic human needs}. The list of basic human needs for which I provide evidence includes needs for physical security, positive affect, relatedness, competence, and autonomy. Although this list should be considered incomplete and revisable, pending additional evidence, I shall take pains to argue for each human need on the basis of recent empirical findings from social and cultural psychology.
version of instrumentalism I shall defend continues to be compatible with the subjectivity of practical reasons.

### 4.4.2 Choosing and pursuing ends

In chapter 1, section 1.1, I summarized instrumentalism as a set of five theses. The thesis defended in this section, which external reasons theorists most often attack, is (I-5):

(I-5) No desired end \( E \) of agent \( S \) is open to criticism from the point of view of practical rationality, except insofar as \( S \) may have a better instrumental reason to pursue some other end, \( E' \).

The charge that external reasons theorists lodge against this thesis is that it’s incompatible with the practical fallibility of ends. If an agent has strong enough desires for very depraved or self-destructive ends, then it’s at least possible for the pursuit of those ends to be an agent’s best reasons for action. And this is unacceptable on its face. Schmidtz, however, suggests that there is in fact no conflict between (I-5) and the fallibility of ends. This can be appreciated, he argues, if we distinguish between four different types of ends (Schmidtz, 2001, pp. 238 – 240). Each type of end is associated with a different sense in which an action is a “means” to achieving an end.

In general, the ends of an agent are the goals for the sake of which an agent acts. An action \( A \) is a means to some end \( E \) if \( E \) is a deductive or causal consequence of \( A \).
Among one’s set of ends are what Schmidtz calls *final ends*. A final end is one which an agent desires and pursues for its own sake, as an “end in itself.” For example: if a jogger wants to run two miles just for the sake of itself, then her goal of running two miles is a final end. Jogging two miles is a means to the final end of jogging two miles, of course, since jogging two miles is a deductive consequence of jogging two miles.

By contrast, agents also pursue what Schmidtz calls *instrumental ends*. An instrumental end is an end pursued as a means to achieving some other end. If a jogger wants to run two miles as a means to being healthy, then her goal of running is an instrumental end. Jogging two miles is a means to the end of being healthy, this time because jogging two miles is a (partial) cause of health.

Third, agents can have *constitutive ends*. A constitutive end is a specification of some further end, such that the fulfillment of the specific end constitutes the fulfillment of the further end. For example, suppose a person wants to get exercise, and in view of this end, she forms the desire to run two miles. In this case, running two miles is not exactly an instrumental end which she pursues as a means to achieving exercise. There is an important difference. Running two miles is a specific form of exercise, among many other possibilities. The jogger could just as easily have chosen to rock climb or play tennis. Since running, rock-climbing, and playing tennis all constitute forms of exercise, they are constitutive ends with respect to the end of getting exercise. Pursuing the constitutive end of running two miles is a means to the end of getting exercise, because
getting exercise is a deductive consequence of running two miles—it’s necessarily that case that, if one runs two miles, then one has gotten exercise thereby.

Instrumentalists usually distinguish between just two types of ends: final and instrumental. In my presentation of the basic claims of instrumentalism from chapter 1 (section 1.1), I restricted the taxonomy of ends to just these two types. Oftentimes in the literature final ends are also called “intrinsic” ends, whereas instrumental ends are called “derivative” ends (Fumerton, 1990, p. 94). But Schmitz adds constitutive ends to the list, as he observes a meaningful difference between one end being a means to achieving a distinct further end, and one end being a specific constituent of a further end for which it is a means.

The distinction between final ends and the other two types of ends is the source of all the resistance to instrumentalism discussed to this point. A final end is neither a constituent of a further end, nor an instrumental end pursued as a means to a further end. In the basic version of instrumentalism I offered in chapter 1, final ends are the terminal points in a chain of practical reasoning. Final ends cannot therefore be subject to criticism from the perspective of instrumental rationality. If the achievement of a final end would maximize the satisfaction of the agent’s desires, then according to instrumentalism the agent’s best reason is to achieve this end, regardless of what that end is. If, for instance, we imagine that Ted Bundy desired the final end of murdering innocent people, and the achievement of this end would lead to the maximal satisfaction of his desires, then Bundy’s best reason would favor going around killing innocent people.
To address this problem, Schmidtz argues that instrumentalism can admit a fourth kind of end which provides a normative basis for criticizing even final ends. These are *maieutic ends*. A maieutic end is an end achieved by a process of *choosing* other ends to pursue (Schmidtz, 2001, p. 139). For Schmidtz, it’s possible to choose a final end as a means to achieving a maieutic end. The final ends people have are not brute facts about our psychologies. They do not “simply assail us,” in Nagel’s phrase. Instead, final ends can be products of choice; one can choose this or that final end through practical deliberation. Schmidtz uses the term “maieutic” in reference to the maieutic method of teaching (i.e., the method of midwifery) attributed to Socrates. The midwife metaphor is apt, since Schmidtz’s idea about the relationship between maieutic ends and other ends is that “we give birth to our final ends in the process of achieving maieutic ends…just as final ends are further ends for the sake of which we pursue instrumental and constitutive ends, maieutic ends are the further ends for the sake of which we choose final ends” (Schmidtz, 2001, p. 239, my emphasis).

Though Schmidtz acknowledges the possibility that maieutic ends can “give birth” to instrumental and constitutive ends as well, he is concerned primarily with explaining how final ends might be chosen as a means to achieving maieutic ends (Schmidtz, 2001, p. 243). Consider, for example, the process one goes through in choosing a career. Schmidtz asks us to imagine Kate, for whom becoming a surgeon is a final end. Kate pursues becoming a surgeon as an “end in itself,” and not as a means to a further end. How did Kate come to have this final end? It might have been that Kate, as a
teenager, did not have a plan for what she wanted to do with her life, but she did want a career of some kind. A few years later, she concludes that becoming a surgeon would be the right career for her. So she resolves to go to medical school and become a surgeon as a final end. The teenage Kate’s desired end of choosing a career can be construed as a maieutic end. This maieutic end is later achieved when Kate chooses the final end of becoming a surgeon. Thus, Kate’s choice of a final end is a means to achieving her maieutic end.

Schmidtz emphasizes that instrumental, final, and constitutive ends are not mutually exclusive. This point turns out to be very important to his full theory, but he does little to elucidate it. Particularly puzzling is Schmidtz’s statement that “[a]n end could be final, pursued for its own sake, and at the same time could be instrumental, pursued as a means to some further end” (Schmidtz, 2001, p. 244). On one interpretation, this sentence sounds like a contradiction in terms. By definition, a final end is an end pursued for its own sake. But if so, how could a final end also be pursued as an instrumental end—i.e., an end pursued for the sake of achieving another end? Aren’t instrumental ends and final ends mutually exclusive by definition? To repel the contradiction charge, Schmidtz would have to suggest that pursuing an end for its own sake doesn’t rule out pursuing that end for the sake of something else. One way to make sense of this is to characterize “pursuing an end” in terms of the representational contents of desires. Let’s say an agent $S$ pursues end $E$ for its own sake just in case the representational content of $S$’s desire for $E$ is the state of the world in which $E$, and only
For example: I jog for its own sake just in case the representational content of my desire to jog is the state of the world in which I jog, and I act on this desire. By contrast, S pursues E for the sake of another end, E’, just in case the representational content of S’s desire for E is the state of the world in which both E and E’ are achieved, and S acts on this desire. For example: I jog for the sake of being healthy just in case the content of my desire to jog is the state of the world in which I jog and I am healthy, and I act on this desire. It’s conceivable that one could have and act on both desires at the same time. I could have a desire to jog for its own sake, and another desire to jog for the sake of being healthy. If I act on both desires, then jogging is simultaneously a final end and an instrumental end for me. Similarly, I jog for the sake of getting some exercise just in case the content of my desire to jog is the state of the world in which I jog and I get some exercise, and I act on this desire. If I act on both this desire and the desire to jog for its own sake, then jogging is both a final end and a constitutive end for me.

Schmidtz anticipates another worry about a potential contradiction in his own terms, this time regarding the relation between maieutic and final ends. If Kate chooses

---

11 Strictly speaking, the representational content of a person’s desire for end E is the proposition representing the state of affairs in which E is the case. But I shall set these technicalities aside.  
12 Also, the representational content of a desire for an instrumental end does not have to be a complete description of what is desired. For instance, Tom may desire to go to university as a means of creating opportunities for himself in the future. Tom’s end of going to university is an instrumental end, since the content of his desire is to go to university and to expand his opportunities. However, Tom does not have to be able to describe in any detail what opportunities he expects to open up for himself—at present, he may have no idea. Similar comments apply to desires for constitutive ends. Jogging two miles could be a constitutive end that Tom pursues for the sake of the further end of getting some exercise. However, Tom may not be able to provide a complete description of that “getting some exercise” involves.
the final end of becoming a surgeon as a means of achieving her maieutic end of settling on a career, is she not by definition pursuing that career as an instrumental end rather than a final end (Schmidtz, 2001, p. 240)? In response, Schmidtz emphasizes a distinction between pursuing an end and choosing an end. He suggests that one’s reasons for choosing an end need not be the same as one’s reasons for pursuing an end. For instance, Kate chooses the career of a surgeon in order to achieve her maieutic end of settling on a career. Achieving the maieutic end was Kate’s reason for choosing a career in surgery. But once she has chosen this particular career, it’s not the case that Kate pursues that career (e.g., by going to medical school) for the sake of settling on a career. Kate’s maieutic end of settling on a career was achieved as soon as she chose to become a surgeon. After that, there is no more maieutic end to be achieved, and Kate can turn to pursuing a surgery career as a final end—i.e., as an end in itself. In sum, a final end cannot by definition be pursued solely as a means to achieving a further end. However, a final end can be chosen as a means to achieving a maieutic end.

An objector might also worry that Kate’s decision to become a surgeon is not a final end, but rather a constitutive end pursued as a means to choosing a career (cf. Schmidtz, 2001, pp. 241 – 42). After all, a surgery career is a specific type of career, and pursuing a surgery career constitutes a specific way of pursuing a career. But here too we must appreciate the distinction between pursuing an end and choosing it. We saw that jogging two miles is a constitutive end pursued as a means to achieving the further end of getting exercise. In order to pursue the end of getting exercise, a specific way of pursuing
that end has to be chosen. Once jogging has been chosen as a way of getting exercise, doing some jogging will constitute a way of getting exercise. Thus, a constitutive end must first be chosen as a prelude to pursuing it as a means to some other end. But it is only in virtue of pursuing a constitutive end (e.g., jogging) that we achieve some further end (e.g., getting exercise). By contrast, a maieutic end is achieved solely by choosing an end, not by pursuing it. Once again, for Kate a surgery career is not pursued as a means to the maieutic end of choosing a career. Rather, a surgery career was chosen as a means to Kate’s maieutic end of choosing a career. Kate’s maieutic end is achieved as soon as she settles on a career. Choosing a career in surgery is not simply a prelude to pursuing her maieutic end; rather, to choose a career in surgery is to achieve the maieutic end of choosing a career. And again, once Kate makes her career choice, she is free to pursue that career as a final end. In sum, a maieutic end is an end achieved by the choice of an end. It is not achieved by the pursuit of an end. Ends chosen in order to achieve maieutic ends are like constitutive ends in that they are specifications of the maieutic end (e.g., a surgery career is a specific kind of career). But whereas the pursuit of a constitutive end constitutes a specific way of pursuing a further end, the choice of an end constitutes a specific way of achieving a maieutic end.

4.4.3 Reflective instrumentalism: leaving no loose ends

Schmidtz’s sophisticated instrumentalism allows for final ends to be subject to rational choice. One final end may be a more optimal means of achieving a maieutic end than another. For instance, Schmidtz points out that “[a] given activity counts as a
prospective career...only if there is a market for that kind of career” (Schmidtz, 2001, p. 246). If Kate has the maieutic end of settling on a career, then this end is more likely to be achieved if she chooses to be a surgeon than a professional shoe shiner. In her cultural milieu, shoe shining may not count as a career, but even if it did, there may be so little demand for professional shoe shining that Kate won’t be able to make a living out of it.

But Schmidtz does not let his account rest by making maieutic ends the “final court of appeals” in assessing the reasons for one’s desired ends. For then the basic objection to instrumentalism would still apply, this time one level up, to maieutic ends. What if someone had the maieutic end of choosing an addiction, or a permanent state of euphoria? It seems very dubious that one could have a reason to choose or pursue these ends. Thus, even maieutic ends can be subject to rational criticism. And yet it’s not clear how a means-ends conception of rationality could provide a basis for evaluating them. The fallibility of ends rears its ugly head once again.

Schmidtz wants to develop an instrumentalist model of rationality which leaves no “loose ends.” The objective is to make every end desired by an agent susceptible to normative evaluation, but in a way that preserves the Instrumentalist Dictum as the master criterion for practical rationality. This can be done by requiring that every desired end of the agent be justified in light of the totality of the agent’s other desired ends. This is exactly what follows from thesis (I-5): if, for any end $E$ desired by an agent $S$, there is no better instrumental reason to choose or pursue another end $E'$, then $S$ can be said to be instrumentally rational (hence practically rational). By acting in accordance with (I-5),
the agent will maximize the satisfaction of his or her desires. Schmidtz calls his model of practical rationality “reflective rationality,” but I prefer to call it *reflective instrumentalism*.\(^\text{13}\)

If final ends can be evaluated with respect to how well they satisfy maieutic ends, how should maieutic ends be evaluated? Schmidtz gives pride of place to *unchosen* ends—ends we simply have, as opposed to ends we acquire by choice (Schmidtz, 2001, p. 245). Unchosen ends, Schmidtz says, “have a certain normative inertia, which means they can be part of a relatively stable frame of reference in terms of which we can evaluate ends we might acquire by choice” (Ibid.). For Schmidtz, unchosen ends are further ends for the sake of which maieutic ends may be rationally chosen or pursued.

By inserting unchosen ends into his framework for reflective instrumentalism, Schmidtz undertakes a descriptive philosophical psychology. He posits two unchosen ends as typical for any human being to desire. The first unchosen end is what Schmidtz calls an “overarching maieutic end.” He describes this end as the end of *having something to live for*. It is “the end of acquiring ends in general…of having one’s life spent on something rather than nothing” (Schmidtz, 2001, p. 246). According to Schmidtz, we have *particular* maieutic ends that apply to particular domains: e.g., the

\(^{13}\) Schmidtz’s terminology differs from mine in another way. He distinguishes his model of “reflective rationality” from what he calls the “instrumental model of rationality” (Schmidtz, 2001, p. 237). In speaking of the “instrumental model,” Schmidtz has in mind a basic version of instrumentalism which only distinguishes between instrumental ends and final ends. On this basic version, final ends are not subject to rational criticism. But I understand “instrumentalism” more broadly as a “means-ends” conception of rationality. Since Schmidtz himself is comfortable characterizing his “reflective rationality” as a means-ends conception of rationality, his view is, in my terms, just a sophisticated form of instrumentalism.
maieutic end of choosing a career, a spouse, an ideal to live by, and so on (Ibid.). The overarching maieutic end of finding something to live for is achieved by choosing at least one particular maieutic end (cf. Schmidtz, 2001, p. 248). Schmidtz tells us that each particular maieutic end is also a constitutive end with respect to the overarching maieutic end. For example: choosing a career is a specific way of choosing something to live for; the choice of the former constitutes the choice of the latter (Schmidtz, 2001, p. 248). Let’s designate the overarching maieutic end as $M$, and each particular maieutic end as $m$.

Schmidtz doesn’t explain why he introduces $M$, but I think the rationale is straightforward. It is a fact of human psychology that people do desire to choose something to live for. Suppose someone passionately desires a particular maieutic end of settling on a career. It’s conceivable that this person also holds this maieutic end as a final end. In that case, there is no further end for the sake of which the agent achieves his end of settling on a career. Accordingly, there would be no further reason for achieving this maieutic end besides the fact that the agent wants to achieve it. However, it makes sense to ask for the reason one might have to settle on a career, and it seems that we can reasonably anticipate a different answer than “I just wanted to choose a career for its own sake.” A plausible alternative answer might be that “I wanted to do something with my life, and pursuing a career is an excellent way to spend one’s time.” Indeed, I think the latter answer is the one most of us would expect to hear. This is because people are not simply possessed by unchosen maieutic ends for choosing ends in particular domains. That young Kate has the particular maieutic end of settling on a career does not seem to
be a brute fact which cannot be explained as being the result of a rational choice. If Kate had grown up in a hunter-gatherer society, where the division of labor is not regimented into specific careers, it seems unlikely that she would ever have formed this maieutic end. Rather, it seems more natural to think that Kate has that maieutic end because its achievement will serve further ends of hers. Kate, like many others human beings, has a desire to spend her life doing something rather than nothing as an unchosen, overarching maieutic end. The particular maieutic ends that Kate adopts in order to serve her overarching maieutic end will be determined by her social environment. Thus, given that Kate inhabits a modern society in which people are encouraged to pursue a career, Kate will likely adopt the maieutic end of settling on a career. If Kate had lived in a hunter-gatherer society, her particular maieutic ends might have ranged over very different domains. Still, no matter what kind of society Kate was reared in, it seems reasonable to expect that she has the end of choosing something to live for. I think this idea is intuitively attractive because it captures a general truth about human psychology: most human beings do, as a matter of fact, want to do something rather than nothing with their lives. (For further explanation of the overarching maieutic end, see chapter 5, section 5.6.2.)

The stronger one’s desire for $M$, it seems, the better one’s reason to choose particular domains in which one can achieve particular maieutic ends. In this way, it seems that even particular maieutic ends can be assessed as means to achieving to a further end—namely, $M$. If I have a strong desire to choose something to live for, but I
fail to form any particular desire to choose something in a particular domain, then I am open to rational criticism. For in that case I would not have maximized the satisfaction of my desires by choosing a domain in which to make choices.

It should also be pointed out that $M$, as Schmidtz describes it, carries no normative freight. Nowhere does he suggest that a person who lacks $M$ somehow fails to be practically rational. Nor does he suggest that there are certain determinate things that an agent must choose to live for, on pain of being practically irrational. Instead, Schmidtz seems to offer the thesis that $M$ is an unchosen end as a descriptive claim about human psychology. $M$ acquires normative significance only insofar as it is the object of desire. For then it will give the agent instrumental reasons to find something to live for.

The second unchosen end which Schmidtz identifies is the end of survival. This end is crucial to his model of reflective rationality, because it enables Schmidtz to “tie up all loose ends.” It makes every end desired by an agent susceptible to practical evaluation by reference to other desired ends. Having identified the overarching maieutic end, Schmidtz introduces survival as the end for which $M$ is a means of achieving. According to Schmidtz, it is a biological fact that organisms have an instinct for survival, since

---

14 Noticeably absent from Schmidtz’s list of unchosen ends is the end of reproduction, or more precisely, the end of propagating one’s genes. Animals of all kinds are endowed with the instinct to bear and rear offspring. Some of them are even disposed to sacrifice their own lives so that their offspring might live. These behaviors are favored by natural selection, since animals that reproduce and sacrifice themselves for their young are more likely to deliver copies of their genes into future generations. I doubt the omission of a reproduction instinct is a mere oversight on Schmidtz’s part. As we saw in our discussion of Foot (section 4.3), human beings are not nearly as compelled by the drive to reproduce as animals seem to be. It may be that a desire for reproduction is not strong or universal enough for Schmidtz’s tastes to count as an unchosen end. I am inclined to agree.

341
organisms with that instinct will be more successful at reproducing themselves than organisms that lack it (Schmidtz, 2001, p. 253). This explains what makes survival an unchosen end. Schmidtz adds that the overarching maieutic end is a means to survival for the reason that “finding things to live for strengthens our will to survive and thereby improves our survival prospects” (Schmidtz, 2001, p. 249). Thus, the overarching maieutic end can be assessed in terms of its effectiveness as a means to survival. If I have a strong desire to find something to live for, then I will be highly motivated to continue doing the things to which I devote myself, and this will require my continued survival. But if I have a weak desire for $M$, there won’t be much that I’m motivated to do, and this can make it less likely for me to survive because I will be only weakly motivated to go on living.

It must be emphasized that Schmidtz does not posit the end of survival as an ultimate criterion for practical rationality. For Schmidtz, survival has no normative significance in and of itself. Rather, whatever normative significance survival does have is derived from the fact that we have instrumental reasons to survive (Schmidtz, 2001, p. 247). According to Schmidtz, it’s simply a fact of human psychology that most human beings desire survival as an end (Schmidtz, 2001, p. 245). None of the ends we choose or pursue could be achieved if we did not live to achieve them. Survival, therefore, is a means to the achievement of any other end one might have, be it instrumental, final, constitutive, or maieutic. Accordingly, survival itself is an instrumental end, and its
normative significance is based on the fact that (most) human beings have instrumental reasons to pursue it.

The end of survival enables Schmidtz to “tie up all loose ends.” Under reflective instrumentalism, we have instrumental reasons to choose or pursue the means to our desired ends. But agents have a diversity of ends related to each other in intricate ways. There are reasons to pursue both constitutive and instrumental ends as means to the achievement of other constitutive or instrumental ends, or as means to the achievement of final ends. There are reasons to pursue final ends for their own sake, and there are reasons to choose final ends as means to the achievement of particular maieutic ends. In addition, there are reasons to choose particular maieutic ends (e.g., the end of choosing a career, choosing a spouse, etc.) as means to achieving the “overarching” maieutic end, $M$. Further, there are reasons to achieve $M$ as a means to achieving survival. And finally, there are reasons to pursue survival as a means to achieving any other end at all.\(^1\)

### 4.4.4 Unification: the fallibility of ends and the subjectivity of reasons

Schmidtz’s model of reflective instrumentalism can already go a long way toward incorporating the practical fallibility of ends. It is particularly effective as a basis for criticizing intuitively imprudent desires. Suppose Kate has the maieutic end of choosing an addiction as a final end. She might choose to become addicted to fast food, gambling, alcohol, or to any one of an assortment of narcotics. Now according to instrumentalism,

\(^1\) For a diagram and summary of this model of reflective instrumentalism, see Schmidtz, 2001, pp. 250 – 51.
just in virtue of the fact that Kate desires this end, she has an instrumental reason to satisfy it. But it hardly follows that this will be her *best* reason. Since addicts characteristically become obsessed with the objects of their addiction, the pursuit of an addiction will tend to crowd out other ends that Kate may desire (e.g., supporting her family, excelling in her career, etc.). Kate’s addiction would even likely frustrate the achievement of the overarching maieutic end of having something to live for. If Kate had no addictions, it’s likely that she could satisfy a greater number of particular maieutic ends for choosing ends within various specific domains. Additionally, certain types of addictions (e.g., to food and drugs) are threatening to one’s very survival. Of course, there are other potentially life-threatening activities, such as race-car driving or mountain-climbing, the choice or pursuit of which are not quite as intuitively contrary to practical reason as choosing an addiction. But nothing about these activities inherently threatens to crowd out the agent’s others desired ends. Also, there is good reason to be skeptical that *most* people you meet will not have enough of a thrill-seeking disposition to support instrumental reasons for driving race cars or climbing K2. Thrill-seekers are, after all, a relatively small segment of the population. So even while we readily accept the some people rationally drive race cars as a pastime, despite the survival risks, we do not think the same of any given person we meet. In the same way that reflective instrumentalism suggests that *most* people will not have reasons to choose an addiction for a final end, it also suggests that most will not have reasons to choose race-car driving.
In sum, although it is not *inconceivable* that Kate’s desires may be such that choosing an addiction would maximize the satisfaction of her desires, the model of reflective instrumentalism suggests that this case is *improbable*. The model makes psychological assumptions about the ends humans generally desire—namely, that people generally have the ends of survival and choosing something to live for. In so doing, it provides a basis from which to see that the choice or pursuit of certain other ends will tend to conflict with those generally desired ends. Provided that these psychological assumptions are true, reflective instrumentalism can account for the intuition that some imprudent ends (like choosing an addiction) can never be rationally chosen or pursued. It can show us that the achievement of some ends is incompatible with the achievement of survival, or a life lived for something rather than nothing, and insofar as nearly all people actually do desire these latter ends, almost all people will have reasons to choose or pursue them.

In sum, reflective instrumentalism consists of a means-ends conception of practical reason, a nuanced taxonomy of ends which does not exclude any end from normative criticism, and a descriptive psychological theory of what human beings generally want. This combination of elements places constraints on the ends which a typical human agent can rationally choose or pursue. Reflective instrumentalism thereby substantiates a *particular interpretation* of the practical fallibility of ends. However, this interpretation is considerably weaker than that favored by the external reasons theorist. The weaker interpretation may be stated as follows:
(FE/W) *The practical fallibility of ends (weak):* Given the facts of human psychology, there are some ends for which it is *improbable* that a human agent has a reason to choose or pursue.

The set of ends to which (FE-W) applies depends on the details of the psychological theory associated with reflective instrumentalism. I will have more to say about this in chapter 5. However, the external reasons theorist might prefer a stronger interpretation of the fallibility of ends:

(Fe/S) *The practical fallibility of ends (strong):* There are some ends for which it is *necessarily* the case that no agent has a reason to choose or pursue.

It must be admitted that reflective instrumentalism conflicts with (FE/S). For according to instrumentalism, reasons are provided by the agent’s desired ends, and there are no necessary limits on the ends which individual humans can desire. If so, there are no necessary limits on the instrumental reasons people can have.

The external reasons theorist might demand not only that the fallibility of ends be substantiated, but also that it be interpreted according to (FE/S). I think this is a mistake, for three reasons. First, to demand that (FE/S) be *justified* only begs the question in favor of external reasons theory. Whether or not there are certain ends which no agent can
possibly have reason to choose or pursue is precisely the issue in dispute. External reasons theorists who hold moral rationalism, for instance, insist that it is impossible for immoral actions to be supported by practical reasons. Instrumentalists deny moral rationalism and (FE/S), because it’s in principle possible for an agent to maximize the satisfaction of his or her desires through vicious acts. To the extent that (FE/S) has intuitive appeal, it is the instrumentalist’s burden to explain away that appeal rather than justify it. I noted in chapter 3 (section 3.1) that instrumentalists can pursue a deflationary critique of moral rationalism. This critique involves an explanation for why we might easily make the mistake of believing moral rationalism is true, when in fact it’s not. The same critique can be lodged against (FE/S). A salient possibility is that we confuse (FE/S) and (FE/W). Perhaps it is only the weak interpretation of the practical fallibility of ends that stands a chance of being true. But insofar as the truth of (FE/W) guides theorists’ intuitions about practical rationality, they might mistake a probabilistic claim for a claim about the impossibility of having reasons to choose or pursue certain ends. It is easy to see how this mistake might be made, particularly if the facts of human psychology are such that hardly any real-life human being would have a reason to pursue some ends.

Second, theories of external reasons which purport to bear out (FE/S) cannot come to terms with the subjectivity of reasons. This was precisely the problem I attributed to Scanlon’s view in section 4.1 and Foot’s recent position in section 4.3. I would suggest a similar problem weighs on other accounts of external reasons: as soon as
the details of such accounts are laid on the table, it becomes very difficult to see how practical reasons can ever be as contingent and idiosyncratic as they sometimes are. The failure of instrumentalism to validate (FE/S) can be compensated by the instrumentalist’s relative success in accounting for the subjectivity of reasons and, at least, (FE/W).

There is a third reason the instrumentalist shouldn’t be expected to provide grounds for (FE/S): it’s probably false anyway. In chapter 1, I offered a non question-begging reconstruction of Williams’s argument for internal reasons. According to that argument, Williams’s explanatory constraint (E-W), together with a psychology of motivation which states that only procedural practical reasoning can motivate agents, implies the conclusion that there are no external reasons. In chapter 2, I gave an argument in favor of (E-W), and in chapter 3, I suggested that there is no behavioral evidence to corroborate the thesis that moral action is motivated by the purely cognitive recognition of moral reasons for action. These lines of reasoning, of course, only re-affirm William’s general conclusion: there are no external reasons. And if that’s correct, (FE/S) entails that no agent can have an internal reason to choose or pursue certain ends. Accordingly, it would have to be the case that procedural principles of practical reasoning necessarily forbid the pursuit of certain ends, whatever the motivations of particular agents may be. This is rather unlikely, since procedural principles are tied to agents’ motivations, and we know that agents’ motivations can vary. Therefore, to the extent that we should hold the instrumentalist responsible for explaining the practical fallibility of ends, we should be satisfied with an account of the weak interpretation of this intuition.
4.5 Conclusion

Though Schmidtz’s account is capable of neutralizing some of the central criticisms of instrumentalism, the model has some gaps. First, it doesn’t do a sufficient job of establishing that human beings do generally desire $M$—the overarching maieutic end of choosing something to live for. For instance, $M$ played an important part in my argument for why it would not be rational for Kate to pick an addiction. But this argument made a crucial assumption: that Kate wants to choose more things to live for, other than her addiction, rather than less. It’s certainly conceivable, however, that Kate could want to choose only an addiction as something to live for, and nothing else. Then her addiction would not conflict with any of her other ends—except perhaps for survival—precisely because she had no other ends. Even if Kate’s addiction was not the fatal kind (e.g., an addiction to shopping), intuitively it would still not be an end she can have reason to choose or pursue. So why would people generally want to choose more things to live for, rather than less? Another important limitation of reflective instrumentalism is that it’s not clear how it could make sense of the intuitive irrationality of immoral desires. As I’ve noted, a major attraction of the concept of external reasons is that it’s a consequence of moral rationalism and moral absolutism. If moral obligations apply independently of any single person’s desires or interests, and obligations necessarily entail reasons, then there must be external reasons. But if the reflective instrumentalist could show that, due to facts of human psychology, most people do have reasons to fulfill their moral obligations, this would go a long way toward explaining
away the appeal of external reasons. Can the reflective instrumentalist plausibly argue that the desires to kill, lie, cheat, or steal either for its own sake, or the sake of further ends, are incompatible with the pursuit of other ends a human agent is likely to have? In chapter 5, I argue that the answer is “yes.”
5. Chapter 5: The rationality of moral compliance

5.0 Introduction

Having laid out a case for instrumentalism about reasons, I now pursue an answer to the “Why be moral?” question. The solution developed here rests on some crucial assumptions that I should make plain immediately. First, I assume an instrumentalist theory of practical rationality. The central thesis I defend in this chapter is that almost all people have instrumental reasons to be moral, almost all the time. All along I’ve interpreted “being moral” as acting morally, and I’ve understood moral action as action that complies with the requirements of morality. Admittedly, I’ve had little to say about morality. Various philosophical and religious traditions offer rich accounts of what morality demands of us. But my approach to the “Why be moral?” question does not require having a complete list of moral rights and wrongs already to hand. My instrumentalism about practical reason enables me to say something about the nature of one’s reasons to be moral, without being committed to any specific moral theory or principle. I take it for granted that most people accept some sentences of the form “It is morally obligatory to X” or “It is morally wrong to Y.” If instrumentalism about reasons is true, then whatever X and Y are, the reasons to do X and refrain from doing Y are that it would satisfy the agent’s desires.

My answer to “Why be moral?” also assumes moral absolutism (cf. chapter 3, section 3.1). According to this thesis, moral obligations apply to people independently of whatever desires they happen to have. In addition, it asserts that people cannot escape
blameworthiness by, say, having a strong desire to do what they are blamed for doing. Thus, assuming that moral absolutism is true, if Ted Bundy has a duty not to murder innocent people, then he has that duty regardless of the fact that he wants to murder. Likewise, Ted Bundy cannot escape blameworthiness for murdering people just because he doesn’t want to refrain from doing it.

The foregoing two assumptions—instrumentalism and moral absolutism—pose a fundamental difficulty for my answer to the “Why be moral?” question. I already spoke to this difficulty in chapter 3, section 3.1, where I discuss the Ted Bundy Paradox. The worry is that instrumentalism and moral absolutism cannot both be true if moral rationalism is true. Moral rationalism is the thesis that moral obligations necessarily entail reasons for action. To be more precise, rationalism says that necessarily, if agent $S$ is morally obligated to do (or refrain from doing) action $A$, then $S$’s best reason for action favors doing (or refraining from) $A$. According to instrumentalism, reasons for action are constituted by the desires of an agent. But we have to recognize that different people can desire wildly different things, and even a single individual can want different things at different times. In chapter 4 (section 4.1), I called this basic fact the contingency of desire. If instrumentalism is true, then the contingency of desire entails that different people’s reasons for action are as diverse and mutually incompatible as their wants. A case in point is Ted Bundy, who (presumably) acted on intense desires to rape and kill people. Supposing that Bundy maximized the satisfaction of his desires by murdering dozens of innocent people, instrumentalism entails that Bundy’s best reason for action
favored murder. Yet many of us will insist that Bundy had a moral obligation to refrain from killing innocent people, no matter what he may have desired. And if moral rationalism is true, then Bundy’s moral obligation entails that his best reason for action favored refraining from killing anyone. Since instrumentalism, moral absolutism, and moral rationalism together imply a contradiction, they cannot all be true. And since I embrace both instrumentalism and moral absolutism, I am committed to denying rationalism.

In chapter 3 (section 3.1), I outlined two possible strategies available to a philosopher, like myself, who wants to argue against moral rationalism while remaining committed to instrumentalism and moral absolutism. The first strategy is the direct critique of moral rationalism, which consists of recruiting normative intuitions that throw moral rationalism into doubt. This was Philippa Foot’s achievement in “Morality as a System of Hypothetical Imperatives” (1972). But I will carry out another line of attack, which in chapter 3 I called the deflationary critique of moral rationalism. This strategy aims to explain away the intuitive appeal of moral rationalism. Undoubtedly, moral rationalism seems to be favored by some powerful normative intuitions. To some, it sounds utterly contradictory to suggest that Ted Bundy has a moral obligation not to murder people, but could possibly have a reason to do the opposite. This intuition has some genuine force, but I do not think it exclusively supports moral rationalism. Our incredulity at the idea that Ted Bundy could have a duty but no reason not to murder people just might be explained by our intuitive, albeit hazy, grasp of a very different
thesis. According to this alternative, it is *highly unlikely* that Ted Bundy, or most people other than Bundy, would have a reason to murder innocent people. Because this proposition makes a weaker claim than moral rationalism, it could be true even if moral rationalism is false. The goal of the present chapter, then, is to provide independent evidence in support of this weaker alternative to moral rationalism. I shall call this alternative the *best bet thesis*, because it suggests that for any agent whose moral duty is to do \( A \), it is a *best bet* to expect that his or her best reason favors doing \( A \).\(^1\) Stated more formally:

*The Best Bet Thesis*

Probably (though *not necessarily*), if agent \( S \) is morally obligated to do (or refrain from doing) action \( A \), then \( S \)’s best reason for action favors doing (or refraining from) \( A \).

The best bet thesis supports the intuition, mentioned above, that it is highly unlikely for Ted Bundy to have a *duty* not to murder people, and yet no *reason* not to do so. The thesis is compatible with moral absolutism and instrumentalism. Moreover, in this chapter I shall provide an argument from instrumentalism to the best bet thesis. It’s a version of a type of argument I’ve called the instrumental defense of morality (cf. chapter 3, section 3.1). The object of the instrumental defense of morality is to show that for almost all people, moral action is almost always instrumentally rational. Hence, the

\(^{1}\) Gavin Lawrence (1995) attributes this same thesis to Philippa Foot’s early work.
instrumental defense of morality can confirm the best bet thesis. The greatest challenge for philosophers who mount the instrumental defense is to give a plausible account of how it could be true that nearly all people maximize the satisfaction of their desires by acting in conformity with moral requirements. As I will stress below, Hobbes, Hume, and their followers have faced up to this challenge with remarkable erudition.

In section 5.1, I introduce my basic approach. I understand moral action to be action motivated by *moral norms*. Section 5.1 introduces the explanatory concept of a *social norm*, and describes the motivational mechanisms by which norms operate. I suggest that moral norms are a subset of social norms, distinguished solely by the fact that moral norms permit, require, or forbid actions that are morally significant. Section 5.2 characterizes several kinds of situations in which norms can be observed to operate. Section 5.3 reviews evidence suggesting that norms explain behavior in the kinds of situations described in section 5.2. Section 5.4 applies the theory of norms in an attempt to enhance the classic answers to the “Why be moral?” question offered by Hobbes and Hume. In effect, I portray reasons to be moral as instrumental reasons to satisfy one’s self-regarding, other-regarding, and moral desires. Section 5.5 considers the question of whether it’s ever rational to try and change the moral norms that happen to prevail in one’s society. An affirmative answer is defended. Section 5.6 extends the argument of section 5.5. There I suggest that moral norms which conflict with one’s *basic human needs* are especially likely to be overturned.
5.1 Moral norms and being moral

In chapter 3 it was argued that moral action is motivated by instrumental reasons. In other words, people are motivated to act in accordance with moral requirements because it serves their self-regarding, other-regarding, and/or moral desires. The problem of this chapter is to explain how an agent’s desires can come into alignment with the demands of morality. That is, how can it be true that a typical human agent who acts to satisfy his or her strongest desires will fulfill his or her moral obligations thereby? After all, there obviously cannot be a necessary connection between moral duty and the satisfaction of desire. Given moral absolutism, some moral duties arguably apply to all people regardless of their particular desires (for instance, the duty not to kill innocent people). But again, we must remember the contingency of desire—different people can want wildly different things. So how can a moral duty apply to just about everyone, and yet be consonant with most people’s desires? If action in conformity with moral duty tends to serve the desires of most people, there has to be some regular relationship between moral duty and human desire. That relationship, I contend, should be characterized as the power of moral norms to induce conformity with our moral obligations. In my view, moral norms generate instrumental reasons for agents to act morally in any of three ways. First, they create punitive social environments in which moral action would serve people’s self-regarding desires to avoid punishment and be well thought-of. Second, they elicit other-regarding desires—e.g., desires to relieve the

--

2 I distinguished self-regarding, other-regarding, and moral desires in chapter 3 (section 3.3).
suffering of others, to be accepted by others, etc.—in most psychologically normal agents. And third, moral norms are internalized, with the result that agents form moral desires to meet their moral obligations for its own sake. In this section, I introduce the concept of a social norm. I argue that moral norms should be understood as a subset of social norms. In effect, I am proposing that moral motivation should be explained in terms of conformity with social norms.

Social norms are a natural kind which social scientists have long theorized about in order to explain the stability and orderliness of human societies. In her recent book, Cristina Bicchieri issues a powerful definition that captures the conditions under which social norms exist (Bicchieri, 2006, p. 11):

**Conditions for a Social Norm to Exist (Bicchieri, 2006)**

Let $R$ be a behavioral rule for situations of type $C$, where $C$ can be represented as a mixed-motive game. Then $R$ is a social norm in a population $P$ if there exists a sufficiently large subset $P_{cf} \subseteq P$ such that, for each individual $i \in P_{cf}$:

1. **Contingency condition**: $i$ knows that a rule $R$ exists and applies to situations of type $C$;

2. **Conditional preference condition**: $i$ prefers to conform to $R$ in situations of type $C$ on the condition that:

   a. **Empirical expectations condition**: $i$ believes that a sufficiently large subset of $P$ conforms to $R$ in situations of type $C$;
and either

(b) *Normative expectations condition*: \( i \) believes that a sufficiently large subset of \( P \) expects \( i \) to conform to \( R \) in situations of type \( C \);

or

(b') *Normative expectations with sanctions condition*: \( i \) believes that a sufficiently large subset of \( P \) expects \( i \) to conform to \( R \) in situations of type \( C \), prefers \( i \) to conform, and may sanction behavior.

Let a *behavioral rule* be a rule which permits, requires, or forbids a certain type of behavior in situations of type \( C \). A behavioral rule \( R \) is not a social norm unless there are a sufficient number of individuals within a population who would prefer to conform to \( R \), provided they have two sets of expectations. First, they must have the requisite *empirical expectations* consisting of the belief that a sufficiently large proportion of the population will conform to \( R \) in the appropriate situations. Second, they must have *normative expectations* consisting of the belief that a sufficiently large proportion of the population either expects them to conform to \( R \) in the appropriate situation, or prefers them to conform and is willing to either reward conformity or punish deviance.

Before unpacking Bicchieri’s definition, a few examples of social norms will clarify what we are talking about (see Elster, 2007, ch. 22; Elster, 2009). Norms of etiquette are well-documented in all human societies. There are, for instance, social norms which regulate how far away one should stand from other people in various social
situations. A partygoer who stands excessively close to other people is apt to find himself with no one to talk to. Gossip might circulate among the other guests, warning people to avoid the oddball. Another familiar rule of etiquette is one which forbids attending a private party without being invited. Elster tells the story of a young officer of pre-revolutionary France, wealthy but not noble, who tried to crash a ball at Versailles. The officer was so harshly ridiculed that upon his return to Paris, he killed himself (Elster, 2007, p. 364). Psychologist Stanley Milgram famously tested a “finders-keepers” norm regulating the allocation of seats in the New York subway system. In the earliest version of the experiment, Milgram told his graduate students to board a crowded subway car and ask someone for their seat. Most of the students recoiled at the idea. When a frustrated Milgram went out to try the experiment for himself, he felt first-hand the severe discomfort of violating such a seemingly insignificant rule. In an interview, Milgram recounted how he felt when finally, after several failed attempts, he asked another subway rider for his seat (Luo, 2004):

Taking the man's seat, I was overwhelmed by the need to behave in a way that would justify my request…My head sank between my knees, and I could feel my face blanching. I was not role-playing. I actually felt as if I were going to perish.

Milgram would eventually persuade ten of his graduate students to complete the experiment. The investigators reported similar experiences of distress while carrying out the experiment.

These examples of etiquette norms illustrate the phenomena that Bicchieri’s definition captures. Social norms are often unspoken rules that people expect one another
to follow. Moreover, norms are enforced by sanctions, which may be positive or negative (Bicchieri, 2006, p. 15). Negative sanctions are punishments of one kind or another.

Observers of norm violations usually feel incited to punish the transgressor. One important type of punishment is ostracism, or the avoidance of future interactions with the transgressor. Keeping one’s distance from the partygoer who stands too close is a relatively mild, yet no less effective mode of ostracism. Another form of punishment is the public expression of contempt, as we saw with the young Parisian officer. For most human beings, contempt can induce intense misery in the target. John Adams wrote that “The desire of esteem is as real a want of nature as hunger; and the neglect and contempt of the world as severe a pain as gout and stone” (Elster, 2009, p. 201). Elster takes the primary emotional effect of contempt on the target to be shame. Thus the young Parisian officer was driven to suicide by the humiliation he suffered. In less extreme cases, shamed norm-violators often try to end their disparagement by withdrawing from the scene or going into hiding. However, shame need not be induced by the express contempt of others. Rather, shame can be generated merely by the awareness that one is being observed violating a social norm. It didn’t take a sneer or a verbal rebuke to put Milgram and his students in a state of mortification. Before they could even bring themselves to breach the subway-seating norm, they were thwarted by their own feelings of embarrassment. All told, the punishments that come with norm-transgression need not take the form of material deprivation or physical injury. Verbal slights, body language and the like can suffice.
Furthermore, punishments can be either *external* and *internal* to the norm-violator. *External punishments* are imposed by others—usually the observers of the violation themselves. *Internal punishments* have their source in the violator’s own psyche. They primarily include guilt, shame and other disturbing emotions brought on by the belief that one has defied social expectations. Positive sanctions, on the other hand, receive relatively less attention in norms theory. But they play just as important a role in explaining norm-compliance as punishment. Positive sanctions supporting norm-conformity include the satisfaction of other-regarding and moral desires, such as the desire to please others or to be fair in one’s dealings. They can also include the satisfaction of being praised by one’s peers.

Now I clarify the details of Bicchieri’s definition. The definition allows for a social norm to exist even if it’s not actually *followed*. Bicchieri calls the members of the subset $P_{cf} \subseteq P$ the *conditional followers* of a social norm $R$. She denotes the subset $P_f \subseteq P_{cf}$ as the *followers* of $R$. Accordingly, a social norm $R$ is *followed* if there are sufficiently many followers of $R$ such that, for each individual $i \in P_f$, conditions 2(a) and either 2(b) or 2(b’) are met for $i$. As a result, $i$ will prefer to conform to $R$ in situations of type $C$. So apart from being followed, a social norm exists in a population as long as there are enough conditional followers who *would* prefer to follow it, *if* they had the appropriate empirical and normative expectations. However, it seems implicit in Bicchieri’s definition that in order for a social norm to be *followed*, the followers’ empirical and normative expectations regarding the norm must be *justified*. For if norm-followers
realize that their expectations unjustified, they may cease to follow $R$ because there is no good reason to believe that enough people expect them to follow $R$.

In addition, Bicchieri defines norms relative to a population, $P$. This permits different norms to exist in different populations. Usually, the “populations” for which one or more social norm exists are cultures—i.e., human communities individuated by shared beliefs and practices. Indeed, cultures are distinguished in part by the different norms that their respective members follow. As Chandra Sekhar Sripada and Stephen Stich point out, the ethnographic record strongly suggests that social norms exist in all human societies, and that all psychologically normal human beings tend to follow the norms that prevail in their local cultural group (Sripada & Stich, 2006, p. 282, 284).

Bicchieri’s definition allows for a norm to exist in a population, even if not all members of the population $P$ would follow it. It is not even necessary for all members of $P$ to know about the norm. The clause stipulating that the set of conditional followers $P_{cf}$ is a subset of $P$ reflects the fact that norms are sometimes violated when they conflict with an individual’s self-regarding interests. Indeed, it is only reasonable to expect some deviance, since the typical kinds of situations to which social norms apply are mixed-motive games. I will have more to say about these games below (Bicchieri, 2006, p. 12).

Condition 2(a) of Bicchieri’s definition is the empirical expectations condition. It states that a conditional follower will prefer conforming to a social norm in situations of type $C$, provided that he or she expects enough other people in the population to conform as well. Condition 2(b), the normative expectations condition, and condition 2(b’), the
normative expectations with sanctions condition, account for the fact that people think of conformity with a norm not only as a descriptive fact about what people will do in certain situations, but also as an obligation with normative significance. When 2(b) or 2(b′) are met, each conditional follower $i$ of social norm $R$ believes that a sufficiently large subset of $P$ “expects” him or her to conform in situations of type $C$. In this context, the “expectations” of the population are to be understood in the normative sense that people think each conditional follower ought to conform to $R$ in $C$. As we saw in chapter 2 (section 2.2), we can distinguish between two interpretations of “ought”—the “ought” of practical rationality, and the “ought” of moral obligation. Bicchieri illustrates the difference between these two interpretations with reference to the rule of driving on the right side of the road (Bicchieri, 2006, p. 14). On the one hand, if everyone drives on the right side of the road, and one does not want to jeopardize her life, then she likewise “ought” to drive on the right side in the sense that she has a practical reason to do it. On the other hand, if everyone drives on the right side of the road, then violating this rule will likely bring harm to others. Accordingly, one “ought” to drive on the right side of the road in the sense that one has a moral obligation not to harm innocent people. For Bicchieri, the relevant interpretation of “ought” which applies to the normative expectations condition is the “ought” of moral obligation.

Many commentators distinguish social norms from other kinds of norms—in particular, legal norms and moral norms. Norms of any kind are behavioral rules enforced by sanctions. However, the sanctions applied to violations of legal norms are importantly
different from those applied to violations of social norms. Legal norms depend on formal enforcers such as courts and police forces. Further, legal norms are formally codified into laws promulgated by governments. Social norms, on the other hand, are not written into law. They are rather the unwritten, sometimes even unspoken, rules we live by. Additionally, social norms are enforced through informal sanctions which can be imposed by anyone who observes a violation. A person who stands too close at a party cannot be arrested by police, but he can be avoided by other partygoers. In California and several other US states, a carousing husband breaks no laws through his infidelity, but he can be shunned by family and friends. Examples can be multiplied with ease.

In my view, moral norms are a subset of social norms. What distinguishes moral norms from social norms is that they permit, require, or forbid morally significant actions. I think we can make an intuitive distinction between moral and non-moral acts. Standing too close to someone at a party may be awkward behavior. At worst, it may be impolite. But it seems strong to call a person who stands too close immoral. Nevertheless, how close people stand to one another in conversation is regulated by a social norm. It is a social norm, but not a moral norm. By contrast, being unfaithful to one’s spouse is not just rude. It’s immoral. In some places, there are legal norms against it. But even where adultery is not illegal, it is still widely considered immoral, and it is regulated by moral norms. Notwithstanding these points, I suggest that moral norms motivate action through the same psychological mechanisms as social norms do generally. A cheating husband may feel guilt for his misdeeds. Likewise, the person who stands too close might come to
feel guilty for making other partygoers uncomfortable. In both the moral and non-moral cases, the internal sanction of guilt may be effective in producing behavior that is consistent with the prevailing norm.

Contrary to my view, some theorists are inclined to make a motivational distinction between social and moral norms (Bicchieri, 2006, p. 20; Elster, 2009). Instead of being a subset of social norms, moral norms might be an entirely different category of behavioral rules, with different psychological underpinnings. Bicchieri, for instance, suggests that one person’s conformity with moral norms does not depend on the expectation that others will conform (Bicchieri, 2006, pp. 20 – 21). Instead, she maintains that people conform with moral norms because of an unconditional preference to follow them. Bicchieri suggests that even in a Hobbesian war of each against all, where all the rest of humanity is engulfed by violence, most people would still prefer to refrain from the immoral act of taking an innocent life. This unconditional preference might flow from a strong emotional aversion to the very idea of killing someone. If Bicchieri is right, then my attempt to characterize moral motivation as being driven by conformity with social norms will be off the mark. However, it seems to me that there are cases of norms which are moral in the sense that their followers speak of them in moral terms—e.g., “It is my responsibility as a good mother to follow norm R.” At the same time, conformity to the moral norm is still conditional on the conformity of others. People will cease to follow the norm if enough others cease to follow it. Moral norms in favor of female genital cutting (FGC) seem to be cases in point (see section 5.5). In some communities where
FGC is practiced, undergoing the procedure is considered both a moral obligation and a sign of good moral character. However, it’s been observed that when enough people repudiate the practice, it is not very long before the rest of the community abandons it as well.

To summarize, norms motivate conformity because agents prefer to conform with them, on the condition that others conform and expect them to do likewise. Call this the conditional preference mechanism. Moral norms can be construed as a subset of social norms. In the following sections, I will explore in greater depth the relationship between moral action and the conditional preference mechanism. By drawing attention to this mechanism, I aim to clarify a sense in which conformity with social norms is instrumentally rational.

5.2 Moral norms and mixed-motive games

Among philosophers there has been growing recognition of the importance of norms in explaining the emergence of moral rules in human social life (Gibbard, 1990; Nichols, 2004; Sripada and Stich, 2006; Harms and Skyrms, 2008). In particular, the theory of norms offers considerable insight into why people follow moral rules, even at great cost to themselves. It also sheds light on how moral action can be motivated by instrumental reasons for action. In this section, I begin to argue that moral action is motivated by moral norms. Moreover, conformity with moral norms is almost always instrumentally rational. This is because conformity with moral norms serves the self-regarding, other-regarding, or moral desires of nearly all human agents. We can begin to
appreciate how moral norms serve people’s desires by looking at the types of social situations in which norms operate. In so doing, it will help to adopt the basic terminology of game theory. The main point here is that moral norms apply to situations where, in the absence of a norm, instrumentally rational action leads to outcomes that are dis-preferred by all the agents involved.

According to Bicchieri, norms apply to what game theorists call mixed-motive games. These are social interactions where there is a conflict of interest among individual “players,” but also a potential for mutual gain (Bicchieri, 2006, p. 3). In mixed-motive games, different outcomes are considered “best” by the relevant players, but there is at least one second- or third-best outcome in which all players can benefit jointly by acting together (i.e., by cooperating).

Two mixed-motive games are especially salient representations of circumstances where moral norms operate. First, there is the one-shot prisoner’s dilemma. The simplest formulation of this game involves just two players, Row and Column, who each must choose between exactly two actions (or strategies) without knowing the choice of the other player. For instance, neither player will know the other’s choice if each chooses simultaneously. There is no prior communication between the players before choices are made. The game is “played” only once, meaning that the two players will not encounter each other again in the future. This makes the game a one-shot prisoner’s dilemma. The

---

3 In the present chapter, I understand “benefit” in purely in terms of utility or desire-satisfaction. Something is good or beneficial for an agent just in case it increases his utility payoff, i.e. leads to greater overall satisfaction of his desires.
two possible strategies are to cooperate and to defect. Each action is associated with a different payoff, which reflects the utility the player receives by taking the action. Also, each player is assumed to be instrumentally rational, which is to say that each player acts to maximize her or her payoff under constraints. However, an individual player’s payoff depends not only on his own choice of action, but also on the choice of the other player. Figure 1 depicts the payoffs received by each player for all possible combinations of strategy choices.

![Figure 1: The prisoner’s dilemma](image)

The dominant strategy of each player in the prisoner’s dilemma is to defect. This means, for instance, that no matter what Column does, it is always best for Row to defect. If Column chooses cooperate, then Row’s best strategy is to defect for a payoff of 4

---

4 When Row and Column are each assumed to be a single agent, the matrix in Figure 1 represents a prisoner’s dilemma. But when Row is a single player and Column is assumed to be a group of (more than 1) players who each receive the payoffs indicated for Column, then Figure 1 represents a public goods game.
(which is better than the 3 Row would get by choosing to cooperate). On the other hand, if Column chooses defect, then Row’s best strategy is once again to defect for a payoff of 1 (which is better than getting 0 from choosing to cooperate). Since Row gets a higher payoff by choosing to defect no matter what Column does, Row’s dominant strategy is to defect. Moreover, notice that the payoffs in this game are symmetrical. Therefore, by the exact same reasoning we followed to determine Row’s dominant strategy, the symmetry of the payoffs implies that Column’s dominant strategy is also to defect.

When rational players follow their dominant strategies in a prisoner’s dilemma, every player chooses defect. This outcome is the equilibrium of the one-shot prisoner’s dilemma. The equilibrium of a game corresponds to the outcome which would result when instrumentally rational agents play the game. Game theorists have devised an assortment of equilibrium concepts for different games of increasing complexity. But the fundamental idea of an equilibrium is that it’s a situation in which each player’s action is a best reply (or best response) to every other player’s action. An action or strategy A for player S is a best reply to the strategy of another player S' when S gets the largest payoff from choosing A, given that S' chooses strategy A' (Bicchieri, 2006, p. 22; Hargreaves Heap & Varoufakis, 2004, pp. 41 – 42). To simplify: in equilibrium, every player in the game chooses the strategy favored by his or her best instrumental reason, given that other players do the same. Equilibria are said to be self-enforcing because no agent has an incentive to deviate from them. In the equilibrium of the prisoner’s dilemma in figure 1, each player maximizes his or her payoffs by choosing defect, no matter what the other
player does. Therefore, defect is the best reply to any strategy chosen by the other player. Since no player has an incentive to change strategies, the equilibrium where every player defects is self-enforcing.

However, the prisoner’s dilemma equilibrium is not the best outcome for all players. When Row and Column both choose defect, each player only receives a payoff of 1. But it’s easy to see that both players would be better off in the situation where each chooses to cooperate. In that case, each player would receive an equal payoff of 3. In the language of rational choice theory, the equilibrium of the prisoner’s dilemma is not Pareto optimal. Pareto optimality is a criterion used in welfare economics to evaluate the welfare of a society.\(^5\) It will help to define this criterion with some precision (Hargreaves Heap et al., 1992, pp. 343 – 344). Suppose we want to compare two possible states of affairs for a society. Let one state be \(W_1\), and the other \(W_2\). If at least one person in the society prefers \(W_1\) to \(W_2\), and every other person either prefers \(W_1\) to \(W_2\) or is indifferent between them, then \(W_1\) is Pareto-preferred to \(W_2\). Under these assumptions, \(W_2\) is Pareto-inferior to \(W_1\). If everyone in the society is indifferent between \(W_1\) and \(W_2\), then \(W_1\) and \(W_2\) are Pareto-indifferent. If some people in the society prefer \(W_1\) to \(W_2\) but others prefer \(W_2\) to \(W_1\), then the two states of affairs are Pareto-non-comparable. Given

\[\]

\(^5\) Paretian welfare economics rests on three postulates. First, an individual’s preferences are taken to indicate that person’s welfare. Second, the welfare of a society depends only on the welfare of the individuals who make it up. And third, if one person’s welfare increases while no-one else’s decreases, then social welfare increases (Hargreaves Heap et al., 1992, p. 343).
these preliminaries, Pareto optimality can be defined as follows (Hargreaves Heap et al., 1992, p. 343):

**Pareto optimality**

A social state of affairs $W$ is *Pareto-optimal* if and only if there exists no other feasible state of affairs $W'$ which is Pareto-preferred to $W$. Otherwise, $W$ is *Pareto-sub-optimal*.

In other words, $W$ is Pareto-optimal just in case no transition $W'$ would make at least one person better off (in terms of utility) than he or she is in $W$ without making someone else worse off. The equilibrium of our prisoner’s dilemma from Figure 1 is Pareto-sub-optimal because there is another outcome which would be Pareto-preferred: namely, the outcome in which each player chooses to cooperate. Indeed, the situation in which both players choose to cooperate is Pareto-optimal, because there is no other possible outcome which would increase one player’s payoff without decreasing the other’s.

Here is central paradox of the prisoner’s dilemma: when each agent acts rationally to maximize his or her utility, the outcome is Pareto-sub-optimal for all. The outcome in which both players cooperate is Pareto-preferred—that is, every player is better off in this outcome to the one where all players defect. However, there is a conflict of interest between the two players, because the *best* outcome for each is to defect while the other
player cooperates. If only there was some way to make it rational for each player to reap the benefits of universal cooperation. This, we shall see, is where norms come into play.

It should be stressed that the prisoner’s dilemma is not some contrived scenario. In fact, it has been studied as an apt representation of a broad range of social interactions (for a review, see Hargreaves Heap and Varoufakis, 2004, ch. 5). Particularly relevant to real-life social affairs are prisoner’s dilemmas involving more than two players. A prisoner’s dilemma with \( n \geq 3 \) players is called a *public goods game*. The public goods game is especially pertinent in situations where a group of people, however large, must act collectively to provide for a public good. A public good is equally and totally consumed by a large group of individuals, where no single individual can be excluded from the benefits of consumption (Brennan, pp. 144-45; Landsburg, p. 486).\(^6\) Canonical examples of public goods include national defense, police protection, city parks, streetlights, clean air, ozone preservation, nuclear disarmament, and democratic government. To *free-ride* is to consume a public good without sharing in the costs of supplying it. The crux of the public goods game is that rational agents will always prefer to free-ride. Free-riding in the public goods game corresponds to the defect strategy in the two-person prisoner’s dilemma. Contributing to the public good corresponds to choosing

\(^6\) Economists define public goods in terms of two dimensions. Some define public goods as non-rivalrous, or “joint.” A good is non-rivalrous when an individual can consume the entire good without reducing the consumption of other consumers. In other words, if a non-rivalrous good is consumed by one person, it is available for others to consume at no additional cost. Others define public goods as non-excludable. A good is non-excludable when people who don’t pay anything for it cannot be excluded from consumption. If a non-excludable good is consumed by one person, it thereby becomes available for others to consume. Free-riding results from the non-excludability of public goods.
to cooperate. In most cases a public good can only be supplied through the contributions of many individuals, so that no single individual’s contribution is enough to provide the good. If too many people fail to contribute, then the few steadfast contributors will get no return for their sacrifice. On the other hand, contributing is costly, and yet it has a negligible impact on whether the good is supplied (especially when many people are contributing). If some people do not contribute while sufficiently many others do, the public good will likely be supplied anyway. Thus, rational individuals will determine that contributing is either vain (when too few people contribute) or unnecessary (when enough people contribute), and their dominant strategy will be to free-ride. If all players free-ride, however, the public good will not be provided at all. This is of course a Pareto sub-optimal outcome, since everyone would be better off consuming the public good.

Another mixed-motive game in which social norms operate is the one-shot _ultimatum game_. We already discussed this game in chapter 3 (section 3.4). A typical one-shot ultimatum game involves two players. The first player, the “proposer,” is given an endowment of money, say 10 monetary units (MUs). The proposer then has to offer any whole number of MUs, from 1 to 10, to the second player, who is called the “responder.” The responder, in turn, can either accept or reject this offer. If the responder accepts the offer, the money is shared according to the terms of the offer. But if the responder rejects, neither player receives anything. In a one-shot ultimatum game, the proposer makes only one offer, and the game ends after the responder accepts or rejects the offer. It is normally assumed that monetary payoffs reflect each player’s utility.
payoffs: i.e., the more money one has, the greater his or her utility. According to traditional game theory, the equilibrium of the ultimatum game occurs when the proposer offers the smallest positive amount to the responder, and the responder accepts that amount. Assuming the proposer can only offer whole MUs in the range of 1 to 10, his best strategy is to offer 1. Given this offer, the responder’s best reply is to accept, since getting 1 MU is better than nothing. Here, each player’s strategy is a best reply to the other player’s. The proposer always prefers that the responder accept an offer. For if the responder rejects, both players get nothing. This is the worst outcome for both the proposer and responder. Knowing that the responder is rational, the proposer can expect the responder to accept any offered amount greater than zero (since some money is better than nothing). At the same time, it is rational for the proposer to offer the smallest positive amount possible. Call this amount $min$. If $d$ is the initial endowment of money (also called the “pie”) in the proposer’s possession, then it is rational for the proposer to maximize his payoffs by offering $d - min$. The ultimatum game is another mixed-motive game because a conflict of interest co-exists with a potential for mutual gain. The proposer would like to offer as little money as possible, whereas the responder would like for the proposer to offer as much as possible. On the other hand, there is potential for joint gain. Both players are better off if they can agree on an offer that leaves each with a positive sum of money.

It’s also worth noting that the equilibrium of the ultimatum game is Pareto-optimal. Given that the proposer gets 9 MUs and the responder gets 1 MU, there is no
way to increase either player’s payoff without decreasing the other player’s. In fact, any agreement where the shares of the two parties add up to the original “pie” turns out to be Pareto-optimal. For instance, an even split where the proposer and responder get 5 MUs each is Pareto-optimal, because there is no way to change the arrangement so that one player gets more money without taking away money from the other.

This completes my review of some important types of situations in which moral norms may arise. In the next section, I explore how the presence of a moral norm can sustain moral action in strategic situations that resemble the prisoner’s dilemma and ultimatum game. Such norms can secure altruistic behavior rather than selfishness, fair and equitable agreements rather than exploitative ones, and so on.

5.3 Moral norms make a difference

The aim of this section is to review evidence for the hypothesis that moral norms can alter behavior in situations of the kind identified in section 5.2. The best evidence comes from experiments with variations of the public goods and ultimatum games. This data documents two patterns of behavior which support the hypothesis that people are psychologically disposed to follow norms. The first pattern of behavior is conformity behavior. This behavior is observed when people abide by a norm upon arriving at beliefs about the relevant situation and the expectations of others. Conformity behavior is predicted by norms theory. If an agent believes that she is in a situation, $C$, to which a behavioral rule $R$ applies, and believes that enough other people conform to $R$ in $C$, and believes that other people expect her to conform to $R$ in $C$, then the agent will want to
follow $R$. Whenever an agent forms beliefs about expected behavior in a given situation, we should expect him to behave as he believes he is expected. Such expectation-consistent behavior is conformity behavior.

The second pattern of behavior predicted by norms theory is *altruistic sanctioning*. It is called “altruistic” because the individual who carries out the sanction incurs a cost, but does not benefit from it. Altruistic sanctioning presents itself in two varieties: rewarding and punishment. Altruistic punishment consists of imposing a punishment on the perpetrator of a norm-violation, even when doing so is personally costly. Altruistic rewarding consists of granting benefits to those who display norm-abiding behaviors. A number of researchers have called the psychological disposition to engage in altruistic sanctioning *strong reciprocity* (Fehr & Fischbacher, 2003). Altruistic sanctioning is predicted by norms theory. Clause 2(b′) in Bicchieri’s definition identifies one condition for the preference to follow norms. On this condition, an individual $i$ believes that a sufficiently large proportion of the community prefers $i$ to follow norm $R$, and may sanction behavior. Of course, the belief that others are prepared to sanction $i$’s behavior can be produced if others do in fact sanction his behavior. And if people sanction any behavior of $i$’s that deviates from a norm, it may be inferred that they prefer that $i$ conform to the norm. Finally, in order for conformity and altruistic sanctioning to count as evidence for a disposition to follow social norms, the agent must not expect to benefit from norm-conformity or altruistic sanctioning in the future. This helps to rule out
alternative explanations for these behaviors, such as the expectation of reciprocated benefits in the future or a concern with maintaining a positive reputation.

Altruistic punishment has been documented by Ernst Fehr and colleagues (Fehr & Gächter, 2002; Fehr & Fischbacher, 2003). This behavior has been shown to occur in one-shot ultimatum games and public goods games. For instance, Fehr and Gächter (2002) ran an experiment in which 240 students were divided into groups of four and made to play a series of public goods games. The games involved real monetary stakes, and there were two treatment conditions: punishment and no punishment. Before the game, each member of the group received an endowment of 20 money units (MUs). During the game, each member was permitted to contribute between 0 and 20 of his or her MUs to a group project (i.e., a public good). Subjects were allowed to keep the money they did not contribute to the project. For every MU invested in the project by a group member, all four group members (even the ones who invested nothing) earned an extra 0.4 MUs. Thus, an investor’s return from contributing 1 MU to the group project would be 0.4 MUs, and the group’s total return from that contribution would be 0.4 x 4 = 1.6 MUs. If all four group members invested all 20 of their MUs in the project, the return for each group member would be 0.4 x 20 x 4 = 32 MUs. However, for each group member, the cost of investing 1 MU in the group project was 1 MU, whereas the return for this investment was only 0.4 MUs. Thus, whatever the other players do, it would always be rational for a player to keep all his or her MUs privately. But if no member
contributed any of their MUs, each would only keep his or her initial endowment of 20 MUs. This game clearly has the structure of a public goods game.

Group members were not informed of each other’s identities. Subjects made their contribution decisions simultaneously, and once their decisions were made, they were informed about the contributions of the other group members. Fehr and Gächter applied two treatments to the groups playing this public goods game. In the punishment condition, subjects could punish any of the other group members after they were informed about each other’s investments. A punishment was imposed by assigning between 0 and 10 points to the target member. Each point cost the punished member 3 MUs. Punishments were also costly to the punisher, however. Each point imposed on the punished member cost the punisher 1 MU. In the no-punishment condition, subjects did not have the option of punishing other group members.

Fehr and Gächter repeated this public goods game in each condition for six periods. Group composition was changed from period to period, so that no subject ever met another subject more than once. This rules out the possibility that a subject might punish a group member in order to induce contributions from him or her in the future. In that case, punishment would not be altruistically motivated; it would be motivated by the expectation of benefiting from the target’s future contributions.

Fehr and Gächter’s findings were as follows. Altruistic punishment was frequent, harsh, and directed at below-average contributors to the group project. In the punishment condition, 84.3% of the subjects punished at least once, and 34.3% punished more than
fives times during the six periods. Most (74.2%) acts of punishment were imposed on below-average contributors, and were executed by above-average contributors (Fehr & Gächter, 2002, p. 137). Also, the more a subject’s contribution fell short of the average group investment, the more the subject was punished. Statistical analysis shows that an increase in a group member’s negative deviation from the average contribution by 10 MUs increased the punishment expenditures of other group members by about 6.22 MUs. Thus, in a four-member group, the resulting payoff reduction the free-rider who contributes 10 MUs less than the average was about 18.66 MUs (Fehr & Gächter, 2002, p. 138). Group members were willing to give up a large fraction of their endowments for the purpose of punishing below-average contributors. This was so even though the punisher could not expect to encounter the target again in a future period of the game. Therefore, punishment could not have been motivated by the expectation of a benefit in future interactions. Furthermore, the fact that below-average contributors were so consistently punished—even at considerable cost to the punisher—strongly suggests that participants preferred that the other players contribute about as much as everyone else.

Fehr and Gächter’s second major result is that altruistic punishment induces greater contributions to public goods. The average investment in the punishment condition was higher in every period than the average investment in the no-punishment condition. Additionally, while contributions increased sharply over time in the punishment condition, it declined in the no-punishment condition. In the final period of the punishment condition, 38.9% of the subjects contributed their whole endowment and
77.8% of subjects contributed 15 MUs or more. In the final period of the no-punishment condition, 58.9% of the subjects contributed nothing and 75.6% contributed 5 MUs or less (Fehr & Gächter, 2002, p. 138). Furthermore, Fehr and Gächter discovered that actual punishment, as opposed to the mere knowledge of the possibility of punishment, increased contributions. When a subject was punished before the last period (period 6), that subject raised his or her investment in the next period by an average of 1.62 MUs. Thus, the effects of altruistic punishment were to increase the provision of a public good to be enjoyed by all members of the population, and to deter decisions to withhold contributions (Fehr & Gächter, 2002, pp. 138 – 39). Because punishment was costly, and the costs could not possibly by recouped in future interactions with the punished party, Fehr and Gächter’s experiment suggests that people are indeed disposed to punish altruistically. Moreover, Fehr & Gächter’s observations of altruistic punishment are well-explained by a norm requiring contributions to the public good. Indeed, it’s difficult to explain why such behavior would occur, unless the players involved generally expected and preferred that people conform to a contribution norm.

Altruistic rewarding has been observed in experiments with the sequential prisoner’s dilemma, also known as the trust game. In sequential games with two players, one player (the first-mover) chooses a strategy prior to the other (the second-mover). Sometimes the second-movers are allowed to know what strategy the first-mover has chosen, but sometimes they are not. The game ends as soon as all players choose their strategy, and it is not repeated.
Over the years it’s repeatedly been found that a much higher proportion of players choose to cooperate in one-shot prisoner’s dilemmas than traditional game theory predicts. Traditional theory predicts that no rational agent would cooperate, since defecting is the dominant strategy. However, numerous studies have shown that people choose to cooperate between 30 and 70 percent of the time (Hargreaves Heap & Varoufakis, 2004, p. 181). More surprisingly, in sequential prisoner’s dilemmas most second-movers choose to cooperate even with the knowledge that the first-mover cooperated (Watabe et al., 1996; Hayashi et al., 1999). This is quite unexpected, since the second-mover would get the largest possible payoff by choosing to defect when the first-mover cooperates. Thus, choosing to cooperate when the first-mover chose cooperate is costly for the second-mover.

Universal cooperation in prisoner’s dilemmas has been observed cross-culturally. Watabe et al. (1996) and Hayashi et al. (1999) ran the same one-shot prisoner’s dilemma experiment with two different samples of college students in Japan and in the United States, respectively. In the Watabe et al. study, each subject played a one-shot prisoner’s dilemma with a partner who participated in another room. Each subject was given 500 yen, and then asked to decide whether or not to give 500 yen to his or her partner. When the subject gave 500 yen, the partner received 1000 yen. When the subject did not give the 500 yen, he or she could keep it. The choice of giving 500 yen corresponds to cooperating in the prisoner’s dilemma in figure 1, while the choice of keeping the 500 yen corresponds to choosing defect. Thus, if both participants gave their 500 yen, both
would receive 1000 yen. If neither participant gave 500 yen, they would keep that amount and get nothing more. If one participant gave the 500 yen while the other kept his or her 500 yen, the former player would end up with nothing while the latter would get 1500 yen. Furthermore, participants were assigned one of five experimental treatments. For present purposes, two treatments are worth highlighting. In the *self-first/knowledge* condition, one partner (the first-mover) made his/her decision before the other (the second-mover), making the game a sequential prisoner’s dilemma. But prior to his/her decision, the first-mover was informed that the second-mover would know of his/her decision. In the *other-first/knowledge* condition, a sequential prisoner’s dilemma again applied, but this time the second-mover was informed of the decision that the first-mover had already made. The Hayashi et al. (1999) experiment was identical to that of Watabe et al. (1996), except that in Hayashi’s study the currency used was US dollars.

Both these studies strongly suggest that people are disposed to reward altruistically. In the *other-first/knowledge condition*, a substantial majority of second-movers chose to give their money *knowing that the first-mover had already chosen to give theirs*. 61 percent of the American subjects cooperated in response to cooperation, and 75 percent of the Japanese subjects did the same (Hayashi et al., 1999, p. 38). This behavior is costly. By choosing to give their money in response to cooperation, the American subjects accepted an opportunity cost of $15 – $10 = $5 and the Japanese subjects accepted an opportunity cost of 1500 – 1000 = 500 yen. Essentially, these second-movers passed on a benefit to first-movers who cooperated, even though it was
costly for them. Furthermore, cooperation by second-movers in the other-first/knowledge treatment was conditional upon cooperation by first-movers. When first-movers defected by keeping their money, only 12% of Japanese subjects cooperated and zero American subjects cooperated. These outcomes fit the pattern of altruistic rewarding.

The Watabe and Hayashi studies also suggest that people expect that others will act in accordance with a norm favoring cooperation in a prisoner’s dilemma-type situation. In post-experiment surveys, subjects who played the role of first-mover in the self-first/knowledge condition were asked whether or not their own choice would affect the choice of their partner. A vast majority of subjects in both experiments (84% in the USA and 83% in Japan) said that they felt their partner would change his or her decision if they changed their own decision (Hayashi et al., 1999, p. 39). In other words, first-movers generally felt that they could control their partner’s behavior by their own action. A majority of first-movers in the self-first/knowledge condition chose to cooperate—56% of American subjects, and 83% of Japanese subjects. An act of cooperation would benefit first-movers only if second-movers cooperated in kind. On the assumption that the first-movers were rational, this suggests that first-movers expected second-movers to cooperate in response to their own cooperation. However, this expectation is inconsistent with traditional game theory, since in one-shot prisoner’s dilemmas and trust games each player’s dominant strategy is supposed to be defection. Thus, a rational player would defect, no matter what the other player does. But the expectation of cooperation in response to cooperation can be explained by a norm which requires cooperation. The fact
that this expectation was confirmed by second-movers who did indeed cooperate in response to cooperation provides even further evidence for a cooperative norm.

Conformity behavior has been observed in studies of framing effects in experimental games. According to Bicchieri, social norms apply to specific types of situations (Bicchieri, 2006, pp. 94 – 99). If an agent acts in conformity with a social norm, she first has to believe she is in the kind of situation where the norm applies. But how an agent categorizes a social situation can depend on the way the situation is framed. For Bicchieri, situations are categorized by cues—features of a situation which make a particular norm salient. A single situation can have multiple sets of cues, and different sets of cues can activate different norms. Framing refers to the process whereby an agent’s attention is focused on one set of cues rather than another. Framing effects occur when different norms can apply to a single ambiguous situation, depending on the way the situation is framed by an agent. In that case the agent will conform to whichever norm he or she takes to be most salient. Conforming behavior can be observed as a result of framing effects, because norms theory predicts that norm-abiding action depends in part on the agent’s interpretation of the situation at hand.

For instance, in experiments with one-shot ultimatum games, people’s choices do not usually lead to the equilibrium predicted by traditional game theory. That equilibrium occurs when the proposer offers the smallest nonzero amount possible to the responder, and the responder accepts it. But in the United States and a number of other industrialized countries, the modal and median offers in experiments with one-shot ultimatum games
are 40 to 50% of the original “pie”, and the mean offers are 30 to 40% (Bicchieri, 2006, p. 104). As noted in chapter 3 (section 3.4), responders reject offers below 20% about half of the time (Gintis et al., 2007, p. 608; Sanfey et al., 2003, p. 1755). Participants in these experiments reportedly reject low offers out of anger at what they judge to be an unfair act (Sanfey et al., 2003, p. 1756; cf. Pillutla & Murnaghan, 1996). Thus, a moral norm governing fair interactions seems to apply in the case of the ultimatum game. When that norm is violated by proposers who make very low offers, angry responders altruistically punish them by rejecting the offer.7

However, different results are obtained in ultimatum game experiments where the strategic structure of the game is preserved, but the game itself and roles of the players are framed differently. Bicchieri cites one study by Hoffman et al. (1994) which framed an ultimatum game in four ways along two dimensions (Bicchieri, 2006, pp. 123 – 125). Under the first dimension, the players’ roles were defined either in terms of the standard proposer/responder relationship, or in terms of a relationship of exchange where half of the participants were assigned the role of “seller,” and half were assigned the role of “buyer.” Under the second dimension, the players’ roles were assigned either “randomly” or “by contest.” In the latter case, participants were ranked by successful performance on

7 The claim that a fairness norm operates in ultimatum games is further supported by another experiment by Fehr and Fischbacher (2003, p. 786). In that study, a third-party player was allowed to punish proposers, but at a cost. Although the third party could not personally benefit from the exchange between proposer and responder, Fehr and Fischbacher observed that third parties were willing to altruistically punish proposers who made low offers. 55% of them punished proposers for offers of less than half the original pie. Also, the lower the offer, the more severe the punishment. In addition, Fehr and Fischbacher found that between 70 and 80% of responders expected that proposers will be punished by third parties for unfair offers.
a pop quiz about current events (Bicchieri, 2006, p. 123). The top six ranked participants were assigned the role of proposer or seller. Thus Hoffman et al.’s study applied four experimental treatments: the exchange/contest frame, the exchange/random frame, the standard/contest frame, and the standard/random frame. The standard/random frame represents the standard ultimatum game.

Participants in each condition were asked to play an ultimatum game in which the objective is to divide $10. As usual, one player is to offer the other some share of the endowment, and the other player can accept or reject it. Offers were lower for games played in the exchange frame than in the standard frame, and they were lower for games played in the contest frame than in the random frame. In the standard/random frame, 85% of proposers offered $4 or more to the responder. In the standard/contest frame, only 60% of proposers offered $4 or more. In the exchange/random frame, 50% of sellers offered $4 or more to the buyer. And in the exchange/contest frame, only 45% of sellers offered $4 or more (Bicchieri, 2006, p. 124). Interestingly, rejection rates remained comparable to those observed in the standard ultimatum game. Although all these treatments are strategically identical to the standard ultimatum game, the different framings elicit different behaviors (at least in players given the roles of seller or proposer-by-contest). The differences in behavior can be explained by norms theory as being due to different expectations elicited by each particular frame. A player designated as a seller may consider herself the owner of the $10, and therefore entitled to divide the sum more selfishly. Likewise, a player designated as the proposer or seller by contest may think of
herself as having earned the $10. In short, the exchange and contest frames activate entitlement norms that are associated with different expectations than the standard ultimatum game.

Further evidence for conformity behavior comes from experiments with one-shot prisoner’s dilemmas and public goods games where players are allowed to communicate with one another before making choices. Bicchieri argues that communication enables the players to agree on a shared framing of their situation. This allows the players to focus on a common norm which governs behavior in that situation. Meta-analyses of public good game experiments by Sally (1995) and Mackie (1997) show that communication uniformly increases cooperation rates, relative to public goods experiments in which no communication is allowed.

A particularly illuminating study by Isaac and Walker (1988) bears out these findings. The authors constructed a public goods game played for two periods, with 10 trials per period. The experiment had three conditions: (1) no discussion in either period; (2) no discussion in period one but discussion in period two; (3) discussion in period one but not in period two. In the cooperation conditions, subjects were allowed to discuss the results of previous trials, but they were not allowed to threaten each other or make side-payments (Isaac & Walker, 1988, p. 589 – 590). Isaac and Walker obtained the following results. In condition (1), cooperation in period one started with 50% in the first trial, but declined to 10% by the tenth trial. In period two, cooperation declined from 40% to zero. In condition (2), cooperation in period one decreased from 50% to 10%, but rose in
period two from 60% to 90%. In condition (3), cooperation was close to 100% throughout period one, but dropped to 85% in period two. Isaac and Walker conclude that discussion helped the group of players to understand that contributions on the part of all players led to a preferable outcome for all. It also enabled the group to credibly agree on the contribution levels to be expected by each member. According to Isaac and Walker, much discussion focused on individuals’ specific agreements to contribute to a coordinated group strategy. On the other hand, the decline of cooperation in condition (3) indicates that continued discussion was necessary to maintain overall commitment to the group strategy. These findings are unsurprising from the standpoint of social norms theory. Bicchieri interprets them to suggest that discussion focused subjects on a relevant norm, which then induced empirical and normative expectations for compliance (Bicchieri, 2006, p. 149).

The effect of unanimous promise-making is also predicted by norms theory. A public goods experiment devised by Orbell et al. (1988) found that in groups where there were unanimous promises to cooperate, 84% of players cooperated. However, in groups in which some promised to cooperate but others didn’t, only 58% cooperated (Bicchieri, 2006, pp. 150 – 151). Bicchieri suggests that unanimous promising can provide a basis for forming empirical expectations about compliance with a behavioral rule (Bicchieri, 2006, p. 173). When people promise to conform to a rule, this is a good indicator of their propensity to follow it. And indeed, the experimental evidence strongly suggests that people typically do follow through on their promises. Thus, unanimous promising can
more effectively satisfy the empirical expectations necessary to activate a conditional preference to comply with a norm.

5.4 The instrumental rationality of following norms

By now we have a fairly strong idea of what norms are and what effects they have on human behavior. At this point the all-important task is to understand how conformity with moral norms can be instrumentally rational. I shall argue that compliance with moral norms is instrumentally rational for most people, most of the time, because it serves any of three kinds of desires: self-regarding desires, other-regarding desires, and moral desires. These three types of desires constitute overlapping motivations to abide by moral norms. Any given human being—even the likes of Hobbes’s fool or Hume’s knave—is likely to have these desires. Moreover, assuming an instrumentalist theory of practical reasons, these three sets of desires also constitute overlapping reasons to follow norms. The arguments of this section will emulate those deployed by Hobbes and Hume in their respective treatments of the “Why be moral?” question. In arguing that moral norms tend to advance one’s self-regarding desires, I look to Hobbes’s example. In showing how moral norms satisfy one’s other-regarding and moral desires, I am more indebted to Hume.

8 For background on Hobbes’s fool and Hume’s knave, see the introduction to this dissertation.
5.4.1 Punishment and rule egoism: Hobbes’s reply to the fool

In *Leviathan*, the fool claims that to “keep, or not keep covenants, was not against reason, when it conduced to one’s benefit” (Hobbes, in SL, p. 614). Hobbes’s concept of a covenant is similar to the notion of a moral norm. It is an agreement or contract in which the parties express the intention to act as agreed, and each party expects that the others will carry out their intentions (Darwall, 1995a, p. 55). We’ve seen that a moral norm involves mutual expectations of compliance, although nothing in the definition of a norm requires that agents *express* their intention to follow it. The fool goes on to point out that it is sometimes “not against reason” to violate a covenant, particularly when the other parties have already complied with it (Hobbes, in SL, p. 614; Kavka, 1995b, p. 6). Given Hobbes’s allegiance to an instrumental theory of reasons (see Darwall, 1995a, p. 59), the fool essentially insists that we sometimes have instrumental reasons to violate moral norms, even when sufficiently many others abide by them. Thus, the focal point of the dispute between Hobbes and the fool is over the rationality of what Gregory Kavka calls *offensive* violations of moral rules—i.e., not complying with moral rules when others are complying with them (Kavka, 1985, in SL, p. 162).

Let me summarize Hobbes’s reply to the fool (cf. Kavka, 1985, in SL, p. 162; Kavka, 1995, pp. 7 – 8). The fool maintains that if one can expect to benefit more from unilaterally violating a covenant (i.e., a moral norm) than from abiding by it, it is instrumentally rational to do so. Hobbes denies this, on the grounds that the costs of error—e.g., being caught and punished—far outweigh the potential gains from norm-
violation. In addition, the chances of error are not precisely calculable. Thus, the uncertainty inherent in violating norms makes it instrumentally irrational to do so. Ultimately, Hobbes offers a rule-egoistic answer to the “Why be moral?” question. By conforming to a set of moral rules without exception, the individual will be more successful in furthering her desired ends than she would be by acting otherwise. In this way, Hobbes defends an instrumentalist version of the best-bet thesis.

Hobbes argues for the rationality of abiding by moral norms solely in terms of external sanctions (cf. section 5.1). Recall from section 5.1 that sanctions can be either positive or negative. Positive external sanctions include the granting of rank, esteem, or status to those who are compliant with norms. By contrast, witnesses of a violation might punish the transgressor with social ostracism, gossip, physical injury, or material deprivation. These are negative external sanctions. According to Kavka, Hobbes restricts his account of the rationality of moral compliance to external sanctions because of his commitment to psychological egoism—the thesis that the only type of desires which motivate people are self-regarding desires (Kavka, 1985, in SL, pp. 161 – 162). However, the view that Hobbes espoused psychological egoism is currently in disrepute among Hobbes scholars (Lloyd & Sreedhar, 2008). For present purposes, we can set aside this interpretive problem, and simply recognize that the restriction to external sanctions is an advantage of Hobbes’s reply to the fool. Few philosophers deny that humans have self-regarding desires, although some are skeptical about whether we are moved by other-regarding and moral desires. Hobbes’s answer to the fool stands a good chance of
persuading even these skeptics. For even if psychological egoism is true, external sanctions are capable of impacting self-regarding desires in ways that generate instrumental reasons to comply with moral norms (Kavka, 1985, in SL, p. 162).

Ultimately, however, I think that a Hobbesian restriction to external sanctions is a mistake. In section 5.4.2, I argue that Hobbes’s account of moral compliance rests on assumptions that are unlikely to hold true unless people are motivated by other-regarding and moral desires.

For better or for worse, *internal sanctions* play no part in Hobbes’s account of moral compliance. Kavka defines negative internal sanctions as “guilt feelings and related forms of psychic distress that most of us are subject to feel when we believe we have done wrong” (Kavka, 1985, p. 165). The negative internal sanctions that Kavka seems to have in mind are the self-critical moral emotions, including guilt, shame, and embarrassment (Haidt, 2003; Moll et al., 2008). Positive internal sanctions, according to Kavka, encompass “the agreeable feelings that typically accompany moral action and the realization that one has acted rightly, justly, or benevolently” (Kavka, 1985, in SL, pp. 165 – 166). These presumably include self-praising moral emotions like pride and elevation (Haidt, 2003; Moll et al., 2008). So characterized, the internal sanctions encompass those emotions which induce other-regarding and moral desires. But let’s not forget: Hobbes restricts himself to explaining how following moral rules can serve the self-regarding desires of individual agents. For ease of communication, I will use the terms “self-interest” and “selfish interest” as synonyms for “self-regarding desire.”
The Hobbesian solution to the “Why be moral?” question rests on two pillars. Both are well-captured by Kavka. First, the possibility of bringing morality into the service of self-interest depends on the social environment. In particular, moral action conduces to self-interest only in what Kavka calls *punitive environments* where immoral persons “are sought out, apprehended, and given stiff punishments frequently enough to make immorality a bad prudential risk” (Kavka, 1985, in SL, p. 164). In a punitive social environment, a sufficient proportion of the population punishes norm-violators, and punishment is both severe and likely enough to occur that the prospective benefits of norm-violation are outweighed by the potential costs. In Hobbes’s moral philosophy, the threat of the sovereign’s sanctions make immorality a bad wager. But the theory of norms suggests that a punitive social environment can be established in the absence of government. We saw in section 5.3 that people are willing to incur a cost in order to punish norm-violators while at the same time rewarding those who uphold norms. Furthermore, punishment ensures that violating norms doesn’t pay. For instance, in the public goods experiment devised by Fehr and Gächter, the individual players who reaped the highest income were the ones whose investment in the public good was close to the average investment of the other players (Fehr and Gächter, 2002, p. 138). Recent theoretical work on the strategic structure of moral norms suggests that individuals conform to moral norms because the threat of punishment makes it in their selfish interest to do so (Sripada, 2005; Boyd & Richerson, 1992).
Punishment is a crucial part of Hobbes’s account of the instrumental rationality of moral compliance. As a source of negative external sanctions, it enforces moral norms. If punishment is sufficiently severe and consistently applied to violators of moral norms, they can give instrumental reasons for even the most selfish of agents to comply. Negative external sanctions make this possible, because they have the power to deter even those who do not have other-regarding or moral desires.

The second pillar of the Hobbesian solution to “Why be moral?” is an idea I shall call the *bounded rationality of moral compliance*. Moral philosophers have long appreciated the futility of arguing that following moral rules necessarily maximizes desirable consequences. It is precisely this point that Hobbes concedes when he addresses the fool’s challenge. For Hobbes, covenants are evaluated on the basis of whether they are optimal means to promoting the agent’s survival (Darwall, 1995a, p. 73). Hobbes never denies the possibility that an agent may come across specific instances—Kavka calls them *golden opportunities*—in which she might advance her ends by violating a norm, *and* there is a negligible chance of being punished (Kavka, 1995b, p. 26). Still, Hobbes denies that it is instrumentally rational to act on the policy, “Obey moral rules most of the time, but seize any golden opportunities to violate them.” In particular, Hobbes disputes two of the fool’s claims. The first claim is that it is “not against reason” for an agent to violate a covenant if the agent *believes* it will promote his or her survival. The second of the fool’s claims is that it is not against reason for an agent to violate a covenant when this will *in fact* promote survival (Darwall, 1995a, p. 73). To make sense
of this position, Kavka attributes to Hobbes a “play it safe” argument against the fool. This argument consists of four main ideas. First, practical reasoning is forward looking: the practical rationality of an action is to be assessed in terms of its effects on the agent’s desired ends after the time of action. Second, choices about whether to comply with a covenant are made under conditions of uncertainty. In such conditions, the probabilities of the various outcomes issuing from the available actions are unknown to the agent. (In rational choice theory, uncertainty is distinguished from conditions of risk, where the probabilities that each of the possible outcomes will occur are known to the agent.) Third, it is rational to play it safe under uncertainty, and thereby avoid extremely bad outcomes in hopes of achieving acceptable benefits. Fourth, a policy of pre-commitment—following moral norms without exception—is instrumentally rational in the sense that one is likely to do better overall by rigidly following moral norms, instead of seeking out golden opportunities for beneficial transgression on a case-by-case basis. This is because pre-commitment avoids bad future consequences that the agent will likely fail to anticipate (Kavka, 1985, in SL, p. 162; 1995b, pp. 21 – 22).

Kavka goes on to expand on Hobbes’s own comments in Leviathan by listing the advantages of pre-commitment compared to the case-by-case search for golden opportunities to violate moral norms. First, following moral norms without exception can minimize the costs of decision, including time, effort, and psychological discomfort. Second, pre-commitment secures against costly biases that make the search for golden opportunities to deviate from moral norms too risky to be worthwhile. Human beings
have a well-documented tendency to over-value immediate rewards compared to later
ones, and to over-estimate their knowledge concerning a practical situation. Furthermore,
even if someone successfully benefits from a golden opportunity to violate a norm and
manages to escape punishment, such initial successes might render the agent
overconfident and thus more susceptible to careless mistakes in the future. Thus,
shortsightedness, self-deception, and overconfidence make the pursuit of golden
opportunities more dangerous than the fool is likely to appreciate. The policy of “Obey
moral rules but seize any golden opportunities to violate them” is so error prone, and the
costs of being subject to external sanctions are so great, that even a purely selfish agent
would more often that not maximize his utility by following moral norms without
exception.

Kavka’s interpretation of Hobbes’s answer to the fool appeals to an idea I call the
bounded rationality of moral compliance. Bounded rationality is the fundamental concept
of a flourishing research program in the cognitive and decision sciences. The most
prominent exponents of this program are economist Herbert Simon (1991) and Gerd
Gigerenzer (Todd & Gigerenzer, 2000). These theorists argue that the process of rational
choice and deliberation must be bounded by the computational limits of the human mind,
and by the structure of the environment within which the search for information about
available options is carried out (Todd & Gigerenzer, 2000, p. 730). Kavka’s attempt to
rationalize moral compliance rests on a bounded conception of rationality. On his view,
the human mind is so fraught with biases and cognitive limitations that the search for
golden opportunities to violate moral norms would be irrational for almost any human agent. The irrationality of trying to exploit golden opportunities can be explained in terms of the instrumentalist theory articulated in chapter 1 (section 1.1). By definition, an instrumental reason for action is a reason favoring the action which the agent *justifiably* believes is an optimal means to achieving his or her desired ends. And according to thesis (I-4), a belief that an action is an optimal means to achieving one’s desires must be justified by an “optimal” amount of evidence (or information). But in chapter 1 (section 1.1), it was suggested that an “optimal” amount of evidence is specified in relation to the background beliefs, preferences, and especially, the resources of the agent. The rational agent’s decisions about how he ought to act are accompanied by “shadow decisions” concerning what quantity of resources should be expended on the search for evidence. In section 1.1, I mentioned that the “ideal” stopping point in the search for evidence would be the point at which the marginal benefit of collecting further evidence equals the marginal cost. However, to think that people can precisely balance the marginal costs and benefits of search is a rather utopian expectation. Theorists of bounded rationality have persuasively argued that it is *psychologically impossible* for real-life human beings to identify the point at which the marginal benefits of information search equal the marginal costs. The task is simply beyond human computational ability. As Todd and Gigerenzer note, human minds do not have “unlimited time and knowledge with which to evaluate the costs and benefits of further information search” (Ibid.). Therefore, on the strength of Williams’s explanatory constraint, it follows that human agents cannot have instrumental
reasons to act on the basis of the “optimal” amount evidence as defined by marginal-cost-equals-marginal-benefit rule.

What, then, is the “optimal” amount of evidence needed to justify an action? Whatever it is, bounded rationality theorists insist that it cannot be defined without taking into account what is known about the capacities of the human mind, as well as the structure of the environments in which human agents act (Ibid.). In a rich body of research, Gigerenzer has shown that people make decisions in accordance with fast and frugal heuristics—rules of thumb that allow for decisions to be made within a minimum of time, and on the basis of a minimal search for information (Todd & Gigerenzer, 2000, p. 731; Gigerenzer, 2008, p. 4). Gigerenzer and his colleagues discovered that fast and frugal heuristics often yield better consequences than statistically or theoretically rigorous, but computationally costly decision rules (Todd & Gigerenzer, 2000, p. 731). One family of fast and frugal heuristics consists of rules involving “one-reason decision-making.” An “unboundedly rational agent” would be able to seek out and evaluate alternative courses of action until the marginal benefit of doing so exactly balanced the

---

9 For example, Gigerenzer discusses a recent study which compared different policies of allocating money across investment portfolios (Gigerenzer, 2008, p. 45; DeMiguel, Garlappi, & Uppal, 2007). A dozen of the policies were based on portfolio theory in economics. One of these theoretically complex policies was that of economist Harry Markowitz, who received a Nobel Prize for his research on investment portfolios that minimize risk and maximize return. The thirteenth allocation policy was the simple $1/N$ rule of allocating equal amounts of money to each option. The performance of the thirteen allocation policies was compared in 7 allocation problems, including one problem which consisted in allocating money to the 10 portfolios tracking the sectors comprising the Standard and Poor’s 500 index. Results of the experiment show that the $1/N$ rule typically made greater returns than the theoretically rigorous policies. This remarkable study suggests that the vicissitudes of the stock market are so unpredictable that someone interested in getting a return on her investments would do just as well, if not better, by following the $1/N$ rule than by relying on some of the best theoretical machinery econometrics has to offer (Incidentally, Markowitz relied on the $1/N$ rule for his own retirement investments!)
marginal cost. A boundedly rational agent, by contrast, would follow a much simpler, faster decision procedure. He or she can select just one criterion by which to evaluate the alternatives. If that criterion singles out one (or a few) of the alternatives as the best, then that alternative may be selected. On the other hand, if the criterion does not discriminate among the alternatives, the agent may select a new criterion by which to evaluate alternatives, and so on until a criterion which does discriminate the alternatives is found (Todd & Gigerezer, 2000, p. 733).

There is good evidence that people act on a “one-reason decision-making” rule which dictates imitating successful actors (Richerson & Boyd, 2005, pp. 124 – 125). “Success” is defined in accordance with whichever criterion the agent uses to discriminate alternatives. In one investment experiment designed by Kroll and Levy (1992), for instance, MBA students were asked to allocate money among three possible investments with different expected returns, return variances, and correlations between returns (Kroll and Levy, 1992; also discussed in Henrich & Henrich, 2007, p. 16). After being given this information, participants set up their investment portfolios and received returns in each of 16 rounds. Between rounds, one of two treatments was applied. In the control treatment, the subjects were allowed to adjust their allocations among investment portfolios, but they did not know about the returns of the other participants. The only information the control subjects had to go on was the data about expected returns, variances, and correlations. In the “imitation-possible” treatment, subjects were given the same information as control subjects, and they were allowed to adjust their portfolios.
But in addition, they were informed of the allocations and performances of all the other participants. Kroll and Levy found that participants in the imitation-possible treatment displayed a strong tendency to imitate those who got the greatest returns. Rather than select their portfolios on the basis of multiple criteria such as expected returns, return variances, and correlations, the subjects opted for the simpler rule of imitating the most successful participants (with “success” being defined as having the highest investment returns). Moreover, the “imitate-the-successful” rule was on average beneficial for those who followed it. Compared to the control group, the allocation decisions of the group of subjects in the imitation-possible condition were on average much closer to the optimal investment allocations predicted by portfolio theory. Fast and frugal heuristics like “imitate the successful” can be advantageous when individually identifying the “best” alternative is error prone or very costly. Sometimes it is not only easier but more beneficial to identify who is successful than the particular strategies that will bring success (Richerson & Boyd, 2005, p. 118, 124).

The Kroll and Levy study shows not only that people do follow fast and frugal heuristics like “imitate the successful,” but also that these heuristics can yield more advantageous results on average than sophisticated, but computationally complex decision procedures. Accordingly, an agent may have an instrumental reason to act on a fast and frugal heuristic. The relatively low expense of employing such a heuristic, together with the fact that a fast and frugal heuristic may actually yield better consequences than more sophisticated decision rules, suggests that employing a fast and
frugal heuristic may be an *optimal* means to one’s desired ends. More than that, a fast and frugal heuristic could even make possible the *maximal* satisfaction of the agent’s desires. And in that case, one has a *best* instrumental reason to use a fast and frugal heuristic.

Thus, a bounded conception of rationality is compatible with instrumentalism. When we take into account the cognitive limitations of the human mind, fast and frugal heuristics may indeed provide human agents with the optimal means by which to maximally satisfy their desires. Bounded rationality theorists tend to eschew the term “optimal” because it is so strongly associated with unbounded models of rationality (Todd & Gigerenzer, 2000, pp. 729 – 730). I however, prefer to retain use of the term while recognizing that various conceptions of optimality exist. Given my commitment to Williams’s explanatory constraint, though, I contend that the conception of optimality relevant to determining an agent’s instrumental reasons must be constrained by the laws governing human psychology. As I noted earlier, if it’s beyond the mental capacities of a human individual to identify the precise point at which the marginal benefits of a search for information equal the marginal costs, then a search of this kind cannot support an instrumental reason for any human agent to act. The findings of the bounded rationality program indicate that the actions which optimally satisfy people’s ends are selected by means of fast and frugal heuristics. If so, then room should be made for a notion of optimality which is “bounded” by considerations of the limits of human cognitive capacities, and this notion of *bounded optimality* should be the one which figures in any view of practical rationality.
Bounded rationality theorists have drawn a connection between the use of fast and frugal heuristics and compliance with social norms. Todd and Gigerenzer suspect that “social norms…can enable fast and frugal social reasoning by obviating cost-benefit calculations and extensive information search” (Todd & Gigerenzer, 2000, p. 740). One way that norms enable fast and frugal decision-making is by providing a social environment in which norm-followers are relatively successful, compared to norm-violators, and therefore worthy of imitation. Robert Boyd and Peter Richerson have devised mathematical models to show how imitating the successful can lead to the emergence of norms which are relatively beneficial to follow (in comparison to non-followers) (Boyd & Richerson, 2000). If the people who follow a norm are more successful on some criterion than those who do not follow it, then they are more likely to be imitated, thus increasing the population of norm-followers.

There are at least two reasons why norm-followers would be more successful on average than norm-violators. First, they do not have to incur the costs of being punished for violating moral norms, and they may even be rewarded for their compliance (e.g., through good reputation). Second, norms sometimes bring about beneficial outcomes only for norm-followers. Recall that in a public goods game, all actors would rather have the public good provided than not provided. In Fehr and Gächter’s (2002) experiment, a norm requiring contribution to the group’s public good seems to enable participants to secure a public good from which all can benefit. The familiar problem with public good games is the temptation to free-ride by consuming the public good without paying for it.
But a norm can eliminate that temptation by making non-contributors susceptible to punishment. The cost of being punished keeps free-riders from benefiting from the public good to the same degree that contributors benefit from it. Thus, those who follow a contribution norm will do better than free-riders because the benefits of the public good are not negated by the costs of being punished. It should come as no surprise, therefore, that in Fehr and Gächter’s experiment the players who earned the highest incomes were those whose individual investments in the public good were close to the average investment of others (Fehr and Gächter, 2002, p. 138). Participants who contributed less than the average received stiff punishments, and consequently were less successful than those who contributed the average amount. On the other hand, players who contributed more than the average received lower payoffs than the contributors who gave just enough to avoid the costs of being punished and still benefit from the public good.

If norm-followers are on average more successful than norm-violators, an agent following the “imitate the successful” heuristic would be inclined to follow moral norms. Moreover, we saw that imitating the successful can be instrumentally rational in some situations. Boyd and Richerson, for instance, argue that in situations where individually assessing the best alternative is difficult or costly, and where the environment changes slowly enough that successful behavior observed in the past will continue to be successful in the future, imitating the successful can yield the best results for the agent (Richerson & Boyd, 2005, p. 118; Boyd & Richerson, 2000). The logic behind Boyd and Richerson’s model offers insight into yet another source of instrumental reasons to follow
moral norms. Hobbes’s fool may wonder about the possibility of benefiting from a moral violation with being punished, but the process of searching for such golden opportunities is shot through with uncertainty, bias, and the danger of catastrophic error. That is, individually assessing golden opportunities is difficult and costly. Furthermore, in a punitive social environment, actors who follow norms will be reasonably successful. They do not have to incur the costs of being punished for violating moral norms, and they may even be rewarded for their compliance. All the fool would have to do to obtain these advantages would be to imitate the norm-followers of his community. Computationally speaking, simply to imitate the successful individuals around him is bound to be less demanding on the fool’s time and resources. Moreover, the fool can probably form a more reliable expectation about the consequences of norm-compliance than he can about the consequences of violation. Assuming the moral norm is commonly followed, the fool would have plenty of role models on which to base a prediction about the benefits he stands to gain from norm-compliance. As long as the moral norm is stable, and not likely to break down anytime soon, conformity can be expected to bring success in the future. By contrast, the fool is unlikely to have many role models who successfully exploit golden opportunities for advantageous norm-violation. And even if he did have such a role model, there is still a strong possibility that he will fail to escape the costs of being punished. Due to shortsightedness, overconfidence, or self-deception, the fool might mistake detectable violations for golden opportunities to violate with impunity. These considerations suggest that in a punitive social environment where moral norms are
enforced by punishment, the fool’s best instrumental reason is to follow moral norms without exception. By employing the fast and frugal heuristic of “imitate the successful,” the fool pursues the boundedly optimal means to his most desired ends.

The key presupposition in the Hobbesian answer to “Why be moral?” is the existence of a punitive social environment. However, Hobbes’s reply to the fool does not address an absolutely crucial issue—namely, how can punitive social environments arise in the first place? This question is particularly thorny when we consider that punishment can be costly for the punisher. Negative external sanctions risk retaliation from the target. The cost of retaliation is particularly high when the punisher is weaker or has less status than the punished. Negative external sanctions are also costly in terms of the effort it takes to monitor the behavior of others. This is likely to prove all the more difficult in a punitive social environment. For in these settings, norm-violators can be expected to develop ever more sophisticated strategies for avoiding detection. Individuals disposed to enforce norms must devise increasingly cunning—and expensive—counter-strategies to see through such machinations. Furthermore, even if someone routinely violates moral norms, it may still be beneficial to interact with them if they are a valuable trading partner or they offer a useful skill.

We thus have no choice but to ask: if punishing is so costly, why do it? In fact, the problem of how to sustain costly punishment of norm-violators has the structure of a public goods game. Punishment can maintain near-universal conformity to moral norms, but those who do not punish benefit just as much from the collective efforts of punishers
as the punishers themselves. This is why the punishment of norm-violators is an altruistic act—individuals cannot expect to benefit from punishing norm-violators anymore than non-punishers, if at all. But then the rationality of punishing is thrown into question. For if enough other people in my community incur the costs of punishing norm-violators, then my personal efforts at punishment will not contribute a significant amount to the general level of norm-compliance (assuming that my community is fairly large). On the other hand, if there are not enough people in my community who punish norm-violators, then my own meager efforts to punish will not have much of an effect in maintaining general norm-compliance. Thus, it is either vain or unnecessary for me to personally take on the costs of punishing, and the rational choice is to refrain from punishing anyone. But the same conclusion applies to everyone else in society. Hence we should expect societies to be dominated by second-order free riders who may themselves comply with moral norms, but fail to punish non-compliers. Hobbes’s reply to the fool only kicks the “Why be moral?” problem up a level to the problem of “Why punish?” A satisfactory answer to “Why be moral?” demands an explanation as to how costly punishment can be sustained. In my view, the most promising explanation can be found in the moral philosophy of David Hume.

5.4.2 Norm-conformity as a virtue: Hume’s response to the sensible knave

Before I offer a complete answer to the punishment problem, I should first point out that there is abundant evidence showing that altruistic punishment does occur. It’s
been widely documented in experimental settings by Fehr, Gächter, and others (for a review, see Gintis et al., 2007). In addition, a considerable amount of field evidence for altruistic punishment has been accumulated. The case studies in Elinor Ostrum’s *Governing the Commons* (1990), in particular, detail the enforcement of moral norms that regulate *common pool resources* such as land, water, and fauna. Usually, the common resource is protected from over-utilization by norms that allocate a limited share to each member of the group of users. Situations involving common pool resources are also classic instances of public goods problems, where all users have an interest in maintaining the resources, but each has a greater interest in taking more than his share. When people use more than their share—for example, when they exceed their allotted ration of water—they are excluded from the group or punished in other ways (Sripada, 2005, p. 777). Ultimately, the question is not whether altruistic punishment—or its counterpart, altruistic rewarding—happens. The problem is to explain why it happens. That explanation will shed light on why it is *rational* to sanction altruistically.

The same motivations that explain altruistic sanctioning also explain why people follow norms at all. To see this, we can no longer assume psychological egoism. It turns out that people are motivated to follow norms not only by a self-regarding desire to avoid punishment, but also by other-regarding desires and moral desires. These additional types of desire can account for altruistic sanctioning. For one thing, if someone harms another person through violating a norm, a third party can be motivated to punish the norm-violator because an empathy-induced desire to help the victim. For another thing,
altruistic sanctioning itself can be enforced by a second-order moral norm which requires the punishment of norm-violators. A second-order norm, \( R_2 \), can be expressed as “Punish anyone who violates rule \( R_1 \),” where \( R_1 \) is a first-order norm. And what would motivate conformity to this second-order norm is a moral desire to follow it as a final end.

David Hume had a deep appreciation for the power of non-selfish desires to motivate moral action. In fact, he anticipated many of the best insights contemporary moral psychology has to offer on conformity to norms. Hume’s response to the sensible knave invokes a psychological mechanism that current commentators know as the internalization of moral norms. I construe internalization as the psychological process through which an individual forms a desire to comply with a given moral norm as a final end.\(^\text{10}\) Recall from chapter 1 (section 1.1) that a final end is an end pursued for its own sake, and not for the sake of any other end. I take it that the process of norm-internalization is responsible for producing moral desires.

Principally in the *Enquiry Concerning the Principles of Morals*, Hume gave pride of price to internalized moral desires. This can be seen in his account of the motivations underlying the virtue of justice. In *A Treatise of Human Nature*, Hume espouses a form of virtue ethics according to which the primary object of moral judgment is a virtue. Virtues, in turn, are defined as character traits that evoke feelings of approval when

\(^{10}\) Sripada and Stich offer the following more lengthy, but I think equivalent, definition of internalization: “According to the internalization hypothesis, individuals exhibit a characteristic style of motivation in which the individual intrinsically values compliance with moral rules even when there is no possibility of sanction from an external source” (Sripada & Stich, 2006, p. 286).
contemplated by an observer in an unbiased way (i.e., from what Hume calls the common point of view) (Cohon, 2004; Cohon, 2006, pp. 256 – 57). Virtuous traits evoke approval in observers through sympathy with those who experience the virtue’s beneficial effects. The moral sentiment of approval for a virtuous trait is what motivates agents to act virtuously. However, Stephen Darwall has noted that the virtue of justice is unlike any other Humean virtue, because it involves an entirely different kind of motivational state. Just persons are distinctive in that they regulate their own conduct not by self-interest, nor by impartial approval for a character trait, but by rules they accept as authoritative (Darwall, 1995a, p. 288). What Hume calls a “convention of justice” arises when members of a society express their common interest in regulating their conduct by certain rules, and actually do so (Darwall, 1995a, p. 292). The most prominent rule of justice in Hume’s ethics is the principle of leaving others in possession of their goods, on the condition that others do likewise. But the motivation for the individual to follow conventions of justice is not founded in the mere empirical expectation that others will regulate their own conduct by them. Rather, it is based on a normative expectation of others, which also applies to oneself, that an individual ought to regulate his or her conduct by rules on the condition that others do the same (Darwall, 1995a, pp. 292 – 293). For Hume, the virtue of justice is ultimately a disposition “to a form of practical reasoning that is realized in the decision to abstain from unjust acts on the ground that it is required by the rules of justice” (Darwall, 1995a, p. 314). The motivation to act according to the virtue of justice is different from the motives for acting on other Humean
virtues (fidelity to promises, benevolence). This special motivation is grounded in what Darwall calls *rule obligation*. Rule obligation consists in the *acceptance* of a rule or norm as a basis for regulating conduct. Rule obligation can be expressed as a judgment, though it will not be a descriptive judgment about what a rule or norm requires. Instead, rule obligation can only be expressed by a *normative* judgment that following the rule is what people *ought* to do (Darwall, 1995a, p. 297). Rule obligation is analogous to what I have been calling *moral desires*: desires to act as morality requires as an end in itself.

Hume’s views on the motivational grounds of justice bear a striking resemblance to the conditional preference for conformity we find in Bicchieri’s definition of a social norm. By that definition, an individual $i$ prefers to conform to behavioral rule $R$ on the condition that enough others conform to $R$ and expect $i$ to conform to $R$. As I’ve noted, the “expectation” for $i$ to follow $R$ is a normative one. That is, for the normative expectations condition to be fulfilled, a sufficient number of people must judge that $i$ *ought to follow* $R$. Moreover, like social norms, Hume’s conventions of justice are also contingent on the empirical expectation that there is a sufficient number of people who follow the convention. There is ample evidence that a conditional preference for conformity is a real motivation in most people. I’ve noted earlier that moral norms explain *altruistic sanctioning* (*rewarding and punishment*) and *conformity behavior*. We saw that most people will cooperate in one-shot prisoner’s dilemmas, given that their partner cooperates first. But if the partner defects first, most will defect in kind. Furthermore, it seems that most people *prefer* the outcome in which all players cooperate.
to the one in which the individual defects in the face of the partner’s cooperation. This is so despite the fact that a player does “best” in purely monetary terms by defecting when the partner cooperates. In post-experiment surveys, Watabe et al. (1996) and Hayashi et al. (1999) asked their subjects the question, “How satisfactory would it be to you if both you and your partner gave 500 yen/$5?” (recall that the Watabe study used yen, and the Hayashi study used dollars). The researchers also asked the question, “How satisfactory would it be to you if your partner gave 500/$5 yen and you did not?” The participants were allowed to rate their satisfaction from 1, meaning “not at all,” to 7, meaning “very much so.” Remarkably, most subjects from both studies (Japanese and American) judged that the result in which both players cooperate would be most satisfactory (Hayashi et al., 1999, p. 30, 37). On the other hand, most participants in both studies judged that not cooperating would yield a more satisfactory result than cooperation when the partner did not cooperate. Since the “self-defect/other-cooperate” outcome is the one that yields the most money, the monetary payoffs must not reflect the players’ actual utilities. Instead, it’s evident that most players in these experiments received greatest satisfaction from cooperating whenever the partner cooperates.

These findings provide compelling indication of a conditional preference for conformity with norms. An important prediction of norms theory is that social norms transform mixed-motive games into coordination games (Bicchieri, 2006, p. 26). In choice-theoretic terms, a player suffers a payoff reduction for violating a social norm. Bicchieri offers a mathematical model to show how this works in the appendix to the first
chapter of her book, and Brian Skyrms gives a useful simplification of it in his recent review (Bicchieri, 2006, pp. 52 – 54; Skyrms, 2008). The seriousness of a norm violation can be measured by the utility deduction suffered by a player (other than the violator) as a result of the violation. We might interpret the seriousness of a norm-violation as the degree to which people are angered or harmed by the offender. A given player’s sensitivity to the violation is measured by a non-negative constant. A player’s utility is his “pre-norm” utility minus the product of his sensitivity and the seriousness of the violation (Skyrms, 2008, p. 465). An individual norm-violator who does not care much about what other people think of him, or about the people harmed by his violation, will have a sensitivity level close to 0. An individual who cares very much about the people affected by his norm-violation will have a sensitivity level closer to 1, and possibly even greater than 1.

We can apply this formula to the prisoner’s dilemma matrix of figure 1. Suppose that in the case of a prisoner’s dilemma, a social norm applies which requires a player to cooperate given that another player cooperates. Each player’s sensitivity to the norm is 0.5. That is, each player cares somewhat about the seriousness of violating the norm, but his utility function discounts the seriousness by one half. Now suppose that the Column player follows the norm, but the Row player defects. The seriousness of the norm violation is 3 – 0 = 3. Column would have received a payoff of 3 if Row had cooperated, but because Row defects, Column receives a payoff of 0 instead. Row’s “pre-norm” payoff in this situation is 4, whereas Column’s is 0. Thus, in the case that Row defects
and Column cooperates, Row’s “post-norm” utility is $4 - 0.5(3) = 2.5$ and Column’s is $0 - 0.5(3) = -1.5$. Since the payoffs of the game in figure 1 are symmetric, the reverse distribution of payoffs occurs when Column defects and Row cooperates. The prisoner’s dilemma from figure 1 is thus transformed into the following game:

<table>
<thead>
<tr>
<th>Column</th>
<th>Cooperate</th>
<th>Defect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row Cooperate</td>
<td>$R = 3, C = 3$</td>
<td>$R = -1.5, C = 2.5$</td>
</tr>
<tr>
<td>Row Defect</td>
<td>$R = 2.5, C = -1.5$</td>
<td>$R = 1, C = 1$</td>
</tr>
</tbody>
</table>

**Figure 2: Norm-transformation of the prisoner’s dilemma in figure 1**

The game is figure 2 is an instance of a *coordination* game in which the best outcome for both players is to “coordinate” their actions in some way. In this simple game, the players coordinate by choosing the same strategy. The game in figure 2 has two equilibria: if one player cooperates, then the other player’s best reply is to cooperate, and if one player defects, then the other’s best reply is to defect. We can think of the conditional preference for conformity as being a function of one’s sensitivity to norm violations and the seriousness of norm-violations. The behavior and post-experiment survey responses of most subjects in the Hayashi and Watabe studies are predicted by the hypothesis that they have a conditional preference for conformity to a cooperative norm.
That is, they cooperate on the condition that others cooperate, but they defect in response to defection.

Norms theory tells us that a conformity preference can arise only if the agent believes that she is expected to follow the relevant norm. We saw in the Isaac and Walker (1998) that players in public goods experiments who are allowed to communicate with one another settle on appropriate contributions to the public good. As a result of this communication, contribution levels increase. This would be predicted by norms theory, given that the participants use the opportunity for communication to agree on expectations about how much each person ought to contribute. Communication would activate a conditional preference to contribute as expected to a norm requiring contribution to the public good. Once activated, this conditional preference can make it rational for each participant to follow the group norm.

Hume’s sensible knave wonders whether, in particular cases, “an act of iniquity or infidelity will make a considerable addition to his fortune, without causing any considerable breach in the social union and confederacy” (Hume, *Enquiry*, 282). Like Hobbes’s fool, the knave voices an aspiration to take advantage of golden opportunities to benefit from violating moral norms, while still enjoying the benefits secured by the conformity of others. Here Hume essentially recognizes the free-rider problem. Moral norms—“conventions” as he calls them—arise because they are mutually advantageous for every member of society. But there may be situations, as in public goods games, where the most profitable outcome is for the individual to violate the norm while
everyone else conforms. One means of counteracting the temptation to transgress norms is the threat of punishment. This is Hobbes’s solution to the problem of securing moral compliance. But unlike Hobbes, Hume does not recommend inflexible conformity to the rules of justice because it is the most advantageous strategy for the individual in the long run. Instead, Hume insists that, as a psychological matter of fact, human beings internalize the moral norms of their community and thereby form moral desires to follow them, so long as others do as well (cf. Darwall, 1995a, p. 310). Once a norm has been internalized, most members of the community will have a conditional preference to conform to it. Such a conditional preference can be described as a moral desire to follow the norm for its own sake. Through internalization, norms enter into our practical deliberations, and give us instrumental reasons to obey them.

There is much we don’t know about how norm-internalization operates. Still, much has been learned already. Sripada and Stich theorize that the psychological mechanisms responsible for internalization are innate. This accounts for the ubiquity of norms in all human societies (Sripada & Stich, 2006, p. 282; 289). Further, Sripada and Stich hypothesize that two main mechanisms subserve internalization. One is responsible for norm acquisition, the other for norm execution (Sripada & Stich, 2006, pp. 289 – 290). The acquisition mechanism identifies the behavioral cues indicating that a norm prevails in a given cultural environment, and passes this information to the execution mechanism. The execution mechanism maintains a database in which the norms identified by the acquisition mechanism are stored. In addition, it generates “intrinsic”
motivations to comply with norms, and to punish norm-violators. I suggest that these compliance and punitive motivations are moral desires to follow norms and to punish violators of norms as final ends. I have already presented a sampling of experimental evidence to show that people do in fact have these desires. But we would also like a more detailed story about how moral desires emerge.

In chapter 3 (section 3.4), I suggested that moral desires are induced by moral emotions. Sripada and Stich, too, hold that the compliance and punitive motivations within the norm-execution mechanism are generated by emotions. What makes an emotion count as a moral emotion is its connection to morally significant judgment and action. What makes judgment or action “morally significant” is determined by the canons of ethics—religious, philosophical, etc. (Ideally, we should prefer moral significance to be defined by the best ethical canon, but I will not here argue whether any such thing exists.) To date, there is much clearer evidence for the involvement of moral emotions in the motivation to punish norm-violations. The most salient instance of a punitive moral emotion is anger or indignation. Fehr and Gächter (2002) found that in experimental one-shot public goods games, free riders trigger strong reactions of anger in other participants, especially in subjects who contribute a lot relative to the free rider. Subjects were asked to respond to various hypothetical stories describing a free-rider who contributes less to the public good than others in the group. Upon reading these stories, subjects reported high levels of anger toward the free-rider. Also, the greater the disparity between the free-rider’s contribution and the average contribution of the group, the
greater the subjects’ self-reported level of anger (Fehr & Gächter, 2002, p. 139). These observations lead Fehr and Gächter to conclude that the altruistic punishment behavior they observed in their public goods experiments was triggered by anger felt by above-average contributors (Ibid.). Furthermore, the anger elicited by free-riding behavior has a moralistic quality. In another public goods experiment allowing communication among group members, Dawes et al. (1977) found that some subjects would invoke moralistic discourse to convince others to contribute. For instance, comments such as “If you defect on the rest of us you’re going to have to live with it the rest of your life” were common (Dawes et al., 1977, p. 7; Frank, 1988, pp. 223 – 225). Lawrence et al. (2002) report that low doses of the dopamine receptor antagonist sulpiride produce selective deficits in several measures of anger, as well as selective deficits in subjects’ willingness to punish others for violations of fairness norms (Sripada and Stich, 2006, p. 294). Thus, anger is a typical emotional response to perceived violations of moral norms. This response moves agents to punish norm violators, even at a cost to themselves. The emotion-induced motivation to punish moral violations can be considered a punitive moral desire—that is, a desire to punish violations of moral norms as a final end.

There is relatively sparse evidence on the emotional bases of motivations to comply with norms. Whereas anger is an other-critical emotion which motivates punishing norm-violators, guilt is a self-critical emotion that motivates compliance with moral norms. Guilt is elicited by the belief that one has caused harm or distress to another person, particularly someone who shares a close personal relationship with the guilty
party. It typically motivates efforts to help the victim of one’s actions or to make up for one’s transgression (Haidt, 2003a, pp. 860 – 861). By contrast, an emotion Jonathan Haidt calls *elevation* is an *other-praising* emotion which appears to make a person want to follow the example of a moral exemplar (Haidt, 2003a; 2003b). Elevation is described as “a distinctive feeling in the chest of warmth and expansion” (Haidt, 2003a, p. 864). It is elicited by witnessing acts of kindness, charity, and self-sacrifice. Further, elevation seems to be a powerful motivator of moral behavior. In one study by Haidt and colleagues, subjects primed to feel elevation via clips about the life of Mother Teresa were more likely than controls to report a desire to help others and become better persons (Haidt, 2003b, p. 282). Haidt received similar reports in a series of interviews with Indian subjects asked to recall a time when they witnessed a good deed done to another person. For instance, one subject had this to say of an act of kindness he once saw: “…another person is doing something for someone, and I wish I could do that. He is doing a good thing, and I feel like this: If I had done this, how much joy I would have felt!” (Haidt, 2003b, p. 283). Thus, elevation seems a likely candidate for a moral emotion that induces moral desires to comply with moral norms. Finally, pride is a *self-praising* emotion which researchers suspect to induce the desire for moral compliance. However, as Tangney and colleagues note, more empirical work exploring the relationship between pride and moral action is needed (Tangney et al., 2007, p. 360).

I’ve been suggesting that we should look to the moral emotions to explain how moral desires are formed through the internalization of moral norms. Moral desires, in
turn, are another source of motivation which makes it instrumentally rational to comply with moral norms as a final end. Hume made a very similar suggestion more than two centuries ago. He tells us that virtuous actions are performed from a sense of obligation (Darwall, 1995a, p. 315). The relevant sense of obligation, here, is rule obligation—the disposition to regulate one’s conduct by the rules one accepts as authoritative. The virtue of justice itself consists in one’s tendency to be moved by rule obligation through practical reasoning (Darwall, 1995a, p. 314). Assuming as many believe that Hume is an instrumentalist, the relevant form of practical reasoning is instrumental reasoning about means to desired ends. Thus, there is a strong resemblance between rule obligation and the internalization of norms. The sensible knave can be understood as a person who is not moved by rule obligation/internalization. Such a person, Hume admits, would lack an important source of reasons to be moral. Speaking about the sensible knave’s proposed strategy of free-riding on others’ compliance with conventions of justice, Hume has this to say (Hume, Enquiry, 283):

If his heart rebel not against such pernicious maxims, if he feel no reluctance to the thoughts of villainy or baseness, he has indeed lost a considerable motive to virtue.

In contrast, a person moved by rule obligation/internalization is sensitive to many more considerations that favor a just life (Ibid.):

Inward peace of mind, consciousness of integrity, a satisfactory review of our own conduct; these are circumstances, very requisite to happiness, and will be cherished and cultivated by every honest man, who feels the importance of them.
With our discussion of Hobbes’s response to the fool, we saw that moral action can serve self-regarding desires to avoid punishment for violating moral norms. Then, turning to Hume, we explored how moral action promotes moral desires to comply with norms for its own sake. But there is a third potential source of instrumental reasons to be moral. These are other-regarding desires. The idea is that other-regarding desires can motivate compliance with moral norms, if the agent judges compliance to be a means of helping or benefiting others. Thus, assuming that moral compliance produces public benefits, compliance may be an instrumental end desired as a means to make those public benefits available to others.

Other-regarding desires can be induced by any of the moral emotions. But empathy is perhaps the best example of a moral emotion capable of inducing other-regarding desires. I gave an analysis of empathy in chapter 3 (section 3.3). The lesson taken from this analysis is that empathy has both an affective component and a cognitive component. It consists of a vicarious experience of another’s suffering or distress, together with a judgment about how the other’s distress is produced in her or her situation. Empathy moves the individual to help, comfort, or alleviate the distress of the other (Haidt, 2003a, p. 862). In principle, empathy can induce either moral desires or other-regarding desires. To see this, recall the case of Roberta, who joins a boycott in protest of the mistreatment of textile workers (cf. chapter 3, section 3.4). If empathy generates an other-regarding desire, the content the desire will solely be the alleviation of the other’s distress, and will not include a judgment about a moral state of affairs. For
example: Roberta joins the boycott out of a desire to alleviate the workers’ suffering. If empathy generates a moral desire, the content of this desire will be a moral state of affairs. For example: Roberta joins the boycott out of a desire to correct what is in her judgment an injustice—namely, the suffering of the textile workers. There is some evidence that empathy induces non-moral other-regarding desires in both humans and animals. Chimpanzees, orangutans, and nonhuman social animals presumably lack the capacity for internalizing moral norms. And yet they have been observed to carry out targeted helping, where one individual’s acts of assistance seem to be grounded in a grasp of the other’s situation. Recall the example of Krom and Jakie, discussed in chapter 3 (section 3.3). These acts of helpfulness may be explained by other-regarding desires, but not moral desires. Moreover, research on human children suggests that emotional reactions to the suffering of others arise in the first year of life. During the second year of life, these reactions begin to motivate attempts to help the sufferer (Haidt, 2003a, p. 861). It is doubtful that the helpful actions of two-year-olds are motivated by moral desires, since current evidence suggests that the competences associated with norm-internalization emerge slightly later (Sripada & Stich, 2006, p. 284). For instance, the ability to reason about normative rules and rule violations appears in children as young as three (Cummins, 1996). These findings suggest that other-regarding desires can serve as a distinct source of motivation for compliance with moral norms. Psychological evidence that humans are motivated by other-regarding desires is vast (for a review, see Sober & Wilson, 1998, chapter 8).
In a similar fashion, Hume appeals to *sympathy* as another motive, distinct from rule obligation, to comply with conventions of justice. For Hume, sympathy ultimately explains why human beings approve of virtue and disapprove of vice (Cohon, 2006, p. 267). When a character trait is advantageous (disagreeable) to its possessor or others, we come to approve (disapprove) of it as a virtue (vice) through sympathy with those affected by it. Upon considering violations of justice by others, Hume suggests, people will typically sympathize with the unease of victims of injustice. Similarly, they will also take sympathetic pleasure when contemplating the social benefits resulting from compliance with justice. These sentiments of approval and disapproval can in turn move people to regulate their own conduct by the rules of justice (Cohon, 2006, p. 263; Darwall, 1995a, p. 314). For insofar as people approve of the capacity for rule-obligation, and disapprove of its absence, they will attempt to cultivate it in themselves. In the terms of my analysis, Humean sympathy could be said to play the same role as empathy in generating other-regarding desires for compliance with moral norms. Aside from the desire to follow moral norms for its own sake, people can conform to moral norms out of a desire to benefit others through moral compliance.

### 5.4.3 Overlapping reasons to be moral

We now have an answer to the problem of sustaining altruistic punishment. As it turns out, we can give the same explanation for altruistic punishment as we just gave for compliance with moral norms in general. And on the basis of this explanation, we can see why in almost all cases it would be instrumentally rational to follow moral norms and
punish others who violate them. This section opened with a Hobbesian account of the instrumental rationality of compliance with moral norms—what Hobbes himself called covenants. The central claim of this account was that it is never instrumentally rational to violate moral norms, because the potential cost of being caught and punished by others is too great. Crucially, the Hobbesian solution presupposed that those who witness norm-violations will punish the transgressor. However, the act of punishing can be costly to the punisher. The costliness of punishing raises a second-order free-rider problem: namely, what makes it rational for individuals to punish norm violations? To answer this question, we had to turn to a Humean philosophical psychology. Whereas Hobbes assumed that human agents are only motivated by self-regarding desires, Hume attributed to people a more diverse repertoire of motivations—in particular, moral desires and other-regarding desires. Moral desires motivate agents to follow norms as a final end. Other-regarding desires motivate agents to follow norms as a means to helping or benefiting others. Both moral desires and other-regarding desires can be induced by moral emotions. If these desires are strong enough, they can reinforce one’s propensity to conform with moral norms, and counteract whatever selfish temptation there may be to violate. More than that, other-regarding and moral desires can make it rational to follow moral norms, when self-regarding desires alone cannot.

If moral and other-regarding desires are common features of human psychology—that is, if most human beings have them—then it will be rational for most people to follow moral norms. Exactly the same comments can apply to altruistic punishment,
assuming there are moral norms that require punishing violators of moral norms. In that case, failing to punish a norm-violator would be a bad wager, because it opens one up to punishment by others. In addition, one is likely to have moral and other-regarding desires that provide instrumental reasons to punish norm-violators, either for its own sake or for the sake of helping the transgressor’s victims. As Sripada points out, the idea of a norm which requires punishment is familiar to moral philosophy. Kant, for instance, argued on the basis of the Categorical Imperative that society has a duty to punish the guilty.

Furthermore, Sripada cites a large body of empirical work in sociology and criminology suggesting that people have powerful retributivist sentiments. That is, many of us think moral offenders should be punished because they ought to suffer, and we think it would be deeply wrong for them to escape with impunity (Sripada, 2005, pp. 785 – 786).

For all that’s been said in favor of moral and other-regarding desires, there are some “never-enders” who would cling to axiom of psychological egoism—i.e., the thesis that most people act only on self-regarding desires for material gains (e.g., money, resources, survival, etc.). Such obstinacy cannot be reasonably maintained, however, in light of the observations presented in section 5.3. In all the main experimental games discussed—one-shot ultimatum games, sequential prisoner’s dilemmas, public goods games, etc.—most participants’ choices departed radically from what purely selfish actors would do. A purely selfish proposer in a one-shot ultimatum game would offer the smallest nonzero amount to the responder. But this is not what most people do: usually, they offer a considerable portion of their endowment. A purely selfish player in a one-
shot public goods game would contribute nothing. Yet many do contribute, and most do so when they are allowed to communicate with other participants. Moreover, these results have been repeated in experiments that control for any possibility that a given participant might gain materially (or avoid a materially adverse consequence) from being generous. For instance, subjects are not allowed to know the personal identities of the other players they interact with, so as to rule out the expectation of future encounters. In some cases the experiments are played with high stakes, involving up to three months’ income (Fehr and Fischbacher, 2003, p. 785). Also, experiments with ultimatum and public goods games have been conducted around the world. Most prominently, Henrich et al. (2004) collected data from 15 small-scale societies in 12 countries on 4 different continents. These societies exhibited a wide variation of economic and cultural practices—from foraging groups to slash-and-burn horticulturalists to nomadic herders, to small-scale agriculturalists. While the researchers observed considerable cross-cultural variation in the way subjects played ultimatum and public goods games, one finding was robust. The model of a purely selfish agent who cares only about monetary payoffs was not supported in any society studied. For instance, in all the societies either the proposer in an ultimatum game gave more than the smallest possible nonzero amount, or the responder rejected a nonzero offer (Gintis et al., 2007, p. 608). Thus, human beings everywhere are motivated by desires which are neither self-regarding nor purely material. The ends we strive for can include what we hope to do for others, and what others expect of us.
5.5 **Unpopular norms: the dark side of conformity**

If any fact is virtually guaranteed by the ethnographic record, it is that norms are universal. They are present in every human society, and almost all human beings reliably internalize the norms of their local cultural group (Sripada & Stich, 2006, p. 282, 284). That norms are universal across societies is practically a sociological “first principle.” However, it is equally beyond question that norms are quite variable. As Stich and Sripada point out, while there is substantial homogeneity in the norms that prevail within human cultural groups, there is also vast diversity in the norms that prevail across groups. Even though some norms can be found in most human societies, such as norms against incest with immediate family members, no known norm prevails in all human societies without exception (Stich & Sripada, 2006, p. 282; Prinz, 2009). For instance, brother-sister marriage (and sex) is known to have occurred in Egypt during the Roman period, and in a number of royal lineages in Hawaii and the Inca empire (Sripada, 2008, p. 332).

So norms are present in all human societies, but the particular contents of norms vary across societies. This raises interesting complications for the “Why be moral?”

---

11 The ambitious comparative study by Henrich and colleagues (2004) attests to both these facts. The researchers found considerable behavioral variation across the fifteen small-scale societies they investigated. Statistical analyses indicated that individual-level economic and demographic factors did not explain this variation, either within or across the cultural groups. Instead, much of the variation was explained by the group’s degree of “market integration” (i.e., how often people buy, sell, or work for a wage) and the extent to which the group engaged in “cooperation in production” (i.e., whether production in the society was usually collective or individual) (Gintis et al., 2007, p. 609). Moreover, the behavior of Henrich’s subjects in the experimental games was consistent with the way people customarily behaved in their particular society. For instance, the Orma of East Africa recognized a similarity between the public goods game and a *harambee*—a contribution that households make when a community decides to construct a road or a school. The Orma called the public goods experiment “the harambee game,” and contributed a generous mean of 58% to the public good. Among the Au and Gnau of Papua New Guinea, many proposers
question, because the extensive cross-cultural diversity in norms includes diversity in moral norms. It’s often the case that most members of one society treat a behavior as permitted or required by a norm, while most people in another society follow a norm which forbids that same behavior. Female genital cutting (FGC) is practiced in many parts of the world. Its practitioners consider it not only prudentially necessary for a girl to marry, but also a sign of good moral character (Mackie, 1996, 2000, 2008). Meanwhile, many people in other parts of the world are revolted by the practice. But in cases such as this, where people in different societies follow conflicting norms, we might ask: Which norms are rational to follow? Is it always instrumentally rational to follow the norms that prevail in one’s community? Or, can it ever be rational to be a moral dissident—i.e., someone who deviates from prevailing moral norms in an attempt to change them? In this section, I argue that in some circumstances it can be rational for most members of a society to challenge the prevailing norms. Some norms, I suggest, are unpopular—they are generally followed within a community, even though most members would prefer to follow a different norm so long as others do as well. I will maintain that such unpopular norms are rationally violated, given certain conditions. In developing this argument I pay

in the ultimatum game offered more than half the original “pie,” and many of these offers were rejected. In the everyday lives of the Au and Gnau, making large gifts is way of bidding for social dominance, and rejection of a generous offer amounts to a rejection of subordination (Ibid.).
special attention to recent efforts by human rights organizations to end female genital cutting (FGC) in African communities.

In previous sections, norms have been cast in a positive light as mechanisms that ensure Pareto-preferred outcomes. For instance, in a public goods dilemma everyone prefers to consume the public good rather than go without it. At the same time, there is an even higher payoff for the individual to consume the public good without contributing anything to it himself. Since it is rational for everyone to succumb to this temptation, no rational agent would contribute anything and the public good would not be provided. But this is a worse outcome for everyone than having the public good. Norms solve this problem by inducing moral or other-regarding desires to contribute to the public good, or to punish non-contributors. Then the interacting group moves to an equilibrium where the public good is secured—a better result for all.

However, there is a dark side to norm-conformity. Sometimes a norm is followed in a population despite the fact that most members of that population prefer, or under ideal conditions would prefer, to follow a different norm. I call such norms unpopular:\(^{12}\)

**Unpopular norms**

Let behavioral rule \( R \) be a norm followed (and perhaps *rationally* followed) by most individuals in \( P \). Suppose \( R' \) is some feasible alternative norm, incompatible with \( R \). Then \( R \) is an *unpopular norm* if, *ceteris paribus*, a majority of the

---

\(^{12}\) Bicchieri (2006, ch. 5) offers a detailed treatment of unpopular norms.
individuals who follow $R$ would prefer to follow $R'$ instead, provided that a sufficient number of individuals in $P$ also follow $R'$.

A few comments on this definition: I shall say two norms are *incompatible* when they require, permit, or forbid mutually exclusive actions in the same situation. For instance, a norm requiring girls of a certain age to undergo genital cutting and another norm forbidding it are incompatible. Further, the definition allows for an unpopular norm to be *rationally* followed. That is, a norm $R$ can be rationally followed by individuals in $P$, while at the same time those individuals would be more satisfied if $R'$ was generally followed instead. This is because the amount of satisfaction one receives from following a norm is contingent on whether (and how many) others follow it. Given that enough people follow norm $R$, individual $i$ might maximize the satisfaction of her desires by following $R$ as well. In that case, $i$ ’s conformity to $R$ will be instrumentally rational. But this does not preclude the possibility that $i$ ’s desires would be even more satisfied by following $R'$ instead of $R$, on the condition that enough other people start following $R'$ instead of $R$.

Next, notice the definition does not guarantee that the more popular norm, $R'$, leads to Pareto-preferred outcomes. While some members of $P$ would get increased utility from a transition to the more popular norm, others might experience a decrease in utility. Even though most people would benefit more from following $R'$, some people might have more to gain from sticking to $R$. In that case, a state of affairs where $R'$ is
followed is not Pareto-preferred. Nevertheless, $R$ can be an unpopular norm relative to $R'$, just in virtue of the fact that most people who follow $R$ would prefer to follow $R'$ (if enough other people follow it as well).\(^\text{13}\)

Recall from section 5.4 that two types of motivations are involved in norm-internalization: punitive motivations and compliance motivations. The punitive motivations that enforce norms make it possible for an unpopular norm to arise. Indeed, punitive motivations make it the case that norms do not have to benefit anyone to be followed. In another paper, Boyd and Richerson (1992) show that altruistic punishment can sustain any arbitrary behavior (Boyd & Richerson 1992; 2005, p. 201; Henrich & Henrich, 2007, pp. 65 - 66). As long as altruistic punishers are sufficiently common in a population, and being punished is more costly than following the norm, it will be rational to follow the norm regardless of what behavior that norm requires. This accounts for why seemingly arbitrary, useless, self-defeating, or even universally harmful norms are sometimes followed. Taboos are prime examples. The anthropologist L.L. Langness reports that the Bena Bena of highland Papua New Guinea refused to eat chicken and eggs because, they said, chickens eat feces. Yet the Bena Bena’s favorite food, pig, also eats feces. The taboo against eating chicken seems not only quite arbitrary, but quite harmful since the Bena Bena were chemically deficient in protein (Edgerton, 1992, p. 56). A similarly bewildering practice was followed by the Ijaw of Nigeria. The Ijaw

\(^{13}\) On the other hand, a norm which produces Pareto-sub-optimal outcomes is unpopular by definition, because a Pareto-sub-optimal state of affairs is one where at least one person can be made better off without making anyone worse off. This is a special case of an unpopular norm, but only a case.
reportedly wanted to have children above all, but they had a custom of killing all newborn twins. When asked why they did this, they answered that it was because “our ancestors did it this way” (Edgerton, 1992, p. 58). Boyd and Richerson’s point is that a social practice does not have to serve any beneficial function for anyone in the relevant communities. All that’s required to sustain such arbitrary behaviors is the preparedness of a sufficient number of people to exact sufficiently costly penalties on deviators.

Compliance motivations also contribute to sustaining arbitrary norms. The conditional preference for conformity motivates agents to follow norms simply for its own sake. This makes some sense of the Ijaw’s rationale for killing twins—“our ancestors did it this way.” The reason to follow the norm is merely that it is or has been followed. This explanation is perfectly understandable if the Ijaw are motivated by a conditional preference to conform. Similarly, an individual Bena Bena can prefer not to eat chicken, if she believes that sufficiently many others both refrain from eating chicken and expect her to do the same. It’s furthermore possible that a higher-order norm for punishing norm-violators maintains an arbitrary practice. Suppose that among the Bena Bena there is a norm requiring the punishment of those who eat chicken. If a sufficient number of Bena Bena prefer to follow this norm, given that others do as well, then the norm against eating chicken will be effectively enforced. Those who might be tempted to eat chicken will be deterred by the enforcers. Thus, assuming that the Bena Bena’s norm

against eating chicken serves no beneficial purpose for any Bena Bena, its existence can still be explained by norms theory. Even if, in fact, the norm against eating chicken allows the Bena Bena to reap some hitherto unknown benefit, we can at least see how in principle it’s possible for an entirely arbitrary norm to emerge in Bena Bena society.

Since any arbitrary rule can in principle become a norm, it’s hardly surprising that norms can be generally followed despite being unpopular. Many social processes can give rise to an unpopular norm. So let us examine one case in detail. Political scientist Gerry Mackie argues in an important series of papers that the practice of female genital cutting (FGC) is maintained by an unpopular norm (1996, 2000; Mackie & Lejeune, 2008). FGC has been documented in 28 countries in Africa, as well as a few countries in Asia and the Middle East. The prevalence of FGC varies among countries, and even among regions within countries, with ethnicity being the key determinant. The World Health Organization estimates that 100 to 140 million women are living with the effects of FGC.¹⁵

Three types of FGC are most common: (1) Clitoridectomy: amputation of part or the whole of the clitoris. (2) Excision: removal of the clitoris and inner vaginal lips. (3) Infibulation: removal of clitoris, some or all of the labia minora, and incisions in the labia to create raw surfaces which are stitched together until they heal into a hood of skin that covers the urethra and most of the vagina (Nussbaum, in Feinberg & Shafer-Landau, ________________

¹⁵ World Health Organization website:
About 85% of women who undergo FGM receive clitoridectomy or excision. Still, in the Sudan, Somalia, and Djibouti, infibulation accounts for 80 to 90% of all procedures (Nussbaum, 1999, in Feinberg & Shafer-Landau, 2008, p. 670).

Mackie himself focuses on infibulation, also known as “pharaonic circumcision.” Depending on local custom, the procedure takes place anywhere from a few days after birth to before the birth of the first child, but usually it is performed around age eight (Mackie, 1996, pp. 1002 – 1003). Except among the affluent, it is done without painkiller or antiseptic. The procedure is usually carried out by women, in private, who hold down the girl and sing and shout to drown out her screams.

Genital cutting as it’s customarily practiced has serious health consequences. Immediate effects include acute pain, hemorrhage, shock, urinary infection, abscess, septicemia, fever, tetanus, and death. Longer-term consequences include pelvic infection, painful intercourse, lack of orgasm, infertility and miscarriage, difficulty in urinating, prolonged or obstructed labor, hemorrhage leading to shock and death, still-born or brain-damaged infants, anxiety, and depression (Mackie, 1996, p. 1003).

Half of the 28 African and Middle Eastern countries where FGC is mainly practiced have introduced legislation forbidding it, and seven more countries have incorporated anti-FGC legislation into their constitutions or criminal laws. However, these laws are often not enforced: prosecutions had been brought forth in only four of the 28 countries. Laws prohibiting FGC have also been adopted in “Western” countries with immigrant communities that continue the practice—Australia, Canada, New Zealand, the
USA, and 13 Western European countries. Apart from these legal measures, a number of non-governmental human rights organizations have initiated campaigns to encourage the voluntary eradication of FGC by practicing communities.

Despite concerted efforts to end or outlaw FGC, the practice has proven resistant to change. Prevalence data collected over the past decade indicates little change in the frequency of FGC. Mackie explains the persistence of FGC by analyzing it as a social norm derived from a *convention*. Conventions are institutions that solve coordination problems. A coordination problem can be represented game-theoretically as a coordination game, such as the following:

<table>
<thead>
<tr>
<th>Column</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
</tr>
<tr>
<td>C2</td>
</tr>
<tr>
<td>Row</td>
</tr>
</tbody>
</table>
| R1     | R = 1, C = 1 | R = 0, C = 0  
| R2     | R = 0, C = 0 | R = 1, C = 1  

**Figure 3: A coordination game with indifferent equilibria**

The payoff matrix in figure 3 represents a coordination game, because the two players—Row and Column—prefer to “coordinate” their strategies. There are two

strategies: strategy 1 and strategy 2. Once again, the players are not allowed to communicate with each other prior to picking a strategy, and they choose strategies simultaneously. If Column chooses strategy 1 (i.e., C1), then Row’s best reply is to choose strategy 1 as well (i.e., R1). And if Column chooses strategy 2 (C2), then Row’s best reply is to choose strategy 2 (R2). The payoffs are symmetric, implying that the best replies for Column are likewise to choose the same strategy as Row. Thus, the game in figure 3 has two equilibria: R1/C1 and R2/C2. In both equilibria, Row and Column choose the same strategy. If they had selected different strategies, both would get nothing. Accordingly, the equilibria of this game are called coordination equilibria. A coordination problem is any game with two or more coordination equilibria. Coordination problems are “solved” when agents act in a way that brings about a coordination equilibrium.

Conventions allow interacting agents who face a coordination problem to converge on a coordination equilibrium. They do so by providing each player with reliable expectations about the actions of others. The canonical example of a coordination problem is that of driving on one side of the road. Whichever side of the road motorists drive on, the important thing is that they all drive on the same side in whatever the direction they’re headed. When all drivers choose to drive on the right, they converge on a coordination equilibrium. But the main difficulty with coordination problems is that the players cannot always form reliable expectations about what the others will do. In terms of the game in figure 3, how can Row form a reliable expectation about which strategy
Column will choose, and \textit{vice versa}? Thomas Schelling (1960) suggests that there is no logical solution to this conundrum. Instead, solutions to coordination problems are achieved by virtue of the fact that some strategies are psychologically \textit{salient} for players. According to Schelling, coordination problems are solved when the strategies that make up a coordination equilibrium are salient to all players. To illustrate, Schelling offers the following example: Suppose you have to meet a friend tomorrow in New York City. But the two of you haven’t agreed on a time or a place. You have no way to communicate with each other, but still you have to meet. You will just have to show up somewhere in New York, at some time, and hope that your friend shows up too. Where should you go, and when? In one experiment, Schelling posed this very problem to a group of subjects. More than 50 percent of them chose the same place: Grand Central Station. And when asked to name a time, almost everyone chose 12 noon (Sugden, 2004, p. 50). This result supports Schelling’s hypothesis that the same strategies are often salient to different people.

Salience is determined by many factors. One important factor is precedence. Suppose two players playing the coordination game in figure 3 arrive at the R1/C1 equilibrium. If they had to play the same game again, the salient choice would be to play the same strategies to arrive at the same equilibrium. Thus, one type of convention is a solution to a coordination problem arrived at due to the precedence of solutions to similar problems faced in the past (Mackie, 1996, p. 1007; see also Bicchieri, 2006, p. 37).
Mackie holds that FGC is a convention in the following sense. Men have an interest in marrying women who will be faithful to them, and who are desirable on a number of other moral, aesthetic, and social criteria. In the communities where FGC occurs, being cut is interpreted as a signal of a girl’s marriage-ability. Accordingly, within an inter-marrying community, it is rational for families to select for their sons wives who have undergone genital cutting. Further, families with daughters expect that families with sons interpret cutting as a sign of spousal quality. In that case, being cut ensures the daughter’s ability to marry. And since girls and their families have both economic and social interests in marriage, subjecting the girl to FGC is a rational course of action (cf. Mackie & LeJeune, 2008, p. 10). Thus, FGC is a convention that facilitates marriage between supposedly high-quality men and women. Mackie’s convention account of FGC is predicated on the assumption that people strongly desire to successfully raise their biological children. Marriage, and FGC, are means which serve that end (Mackie, 1996, p. 1007).

FGC is more than a convention in some communities. Rather, in places where the practice is maintained by what Mackie calls an honor and modesty code, FGC has all the trappings of a moral norm. Under an honor and modesty code, genital cutting is taken to be a sign of virtue. By diminishing female sexual desire, it is thought to indicate chastity and fidelity—traits considered pre-requisite for marriage. In addition, mothers in

---

18 On Bicchieri’s definition, norms apply to mixed-motive games. Conventions are solutions to coordination games. But Bicchieri points out that overtime conventions, too, can become norms when they are augmented by a sanctioning system (Bicchieri, 2006, pp. 38 – 39).
some communities consider finding a husband for their daughters to be among the duties of a good parent (Mackie & LeJeune, 2008, p. 29). Mackie notes that in many areas where FGC is practiced, “marriage is a woman’s primary source of material subsistence beyond early adulthood” (Mackie & LeJeune, 2008, p. 10). Given the importance of genital cutting as a sign of marriage-ability, mothers often look upon FGC as a necessary means to securing a prosperous future for the daughters they so love. Also, Mackie suggests that FGC may also be rooted expectations of reciprocity in marriage. Traditionally it’s thought that the male owes his wife material support during pregnancy and child-rearing, while the female owes her husband a child of his own (Mackie & LeJeune, 2008, p. 31). FGC is a way for a wife to assure fidelity and paternity to her husband. In light of these seemingly moral considerations favoring FGC, as well as the fact that the practice is enforced by informal sanctions, I take it that FGC has the status of a moral norm in many practicing communities.

However, not all rationales for FGC appeal to moral considerations of virtue, reciprocity, and honor. Some are more aesthetic in nature. It’s sometimes believed that FGC confers beauty and cleanliness (Mackie & LeJeune, 2008, p. 10). Still other rationales are dissociated from moral and aesthetic judgments, and instead emphasize the alleged pragmatic advantages of FGC as a way of avoiding untimely pregnancy in girls (Mackie, 2000, p. 269). There are also a number of false factual beliefs to the effect that FGC has positive health consequences. For example, it is said to promote health and fertility, to make child-bearing easier, and to enhance male intercourse. These factual
claims have been thoroughly discredited (Mackie, 1996, p. 1005; Mackie, 2003). Finally, FGC is thought to be prescribed by religion—usually Islam, but sometimes Christianity as well. These allegations are also incorrect (Mackie, 1996, p. 1004).

The rationales for FGC listed above create what Mackie calls a belief trap (Mackie, 1996, p. 1009). They help to maintain an FGC convention by contributing to a sanctioning system which incentivizes people to keep practicing FGC. It is this sanctioning system that gives the FGC convention the characteristics of a norm. External sanctions imposed on a girl or woman who is uncut include stigmatization, social ostracism, and isolation from her peer group. Foregoing FGC can lead to a loss of social standing for the girl and her family, and marrying an uncut woman can similarly lead to a loss of status for a husband and his family (Mackie & LeJeune, 2008, pp. 26 – 27).

Internal sanctions are also important factors in motivating the decision to be cut. Virtually every ethnographic account states that FGC is transmitted and defended by women (Mackie, 1996, p. 1003). Again, mothers sometimes endorse it as part of their responsibility to raise a girl “properly” (Mackie & LeJeune, 2008, p. 29). In Darfur, Sudan, highly educated Arab-Sudanese women tend to view FGC favorably, as a sign of ethnic superiority over groups in the region that do not practice it (Mackie, 2000, p. 268). These observations are hints that women internalize a moral norm that requires genital cutting.

Despite the incidence and persistence of FGC, Mackie insists that it is an unpopular norm. His claim is that those who choose to make their daughters endure genital cutting would prefer to abandon it, if they were informed about its consequences and about possible alternative ways of life, and if enough others abandoned the practice as well. Mackie’s hypothesis is somewhat supported by survey and ethnographic evidence. For instance, FGC persists even among families who agree that it should be abandoned. Furthermore, a 1980 survey of 70 Somali female and 40 male university students revealed that 60 percent of women and 58 percent of the men believed that FGC should be abolished. Nevertheless, 66 percent of women and 50 percent of men planned to cut their daughters (Mackie, 1996, p. 1014). But Mackie takes the primary evidence for the unpopularity of FGC to be the remarkable success of activist organizations in bringing an end to the practice in some African communities (Mackie, 2000). Apparently, these organizations were able to initiate epidemics of abandonment. Mackie cites two advances as contributing to these epidemics (Mackie & LeJeune, 2008, p. 1). The first is Mackie’s own theory of FGC as a social convention, which has informed successful approaches for the eradicating FGC. And the second is the establishment of programs that organize the collective abandonment of FGC in villages through a strategy of community dialogue and human rights education (Ibid.).

The convention account of FGC shows how it can be rational for agents to follow an unpopular norm, even though they would prefer to follow an alternative norm on the condition that a sufficient number of others follow the alternative as well. Mackie compares the FGC convention to a Pareto-sub-optimal equilibrium in a coordination game with ranked equilibria. Here is an illustration:

<table>
<thead>
<tr>
<th>Row</th>
<th>Column</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td></td>
<td>R = 2, C = 2</td>
<td>R = 0, C = 0</td>
</tr>
<tr>
<td>R2</td>
<td></td>
<td>R = 0, C = 0</td>
<td>R = 1, C = 1</td>
</tr>
</tbody>
</table>

**Figure 4: A coordination game with ranked equilibria**

The game in figure 4 is another coordination game with two equilibria: R1/C1 and R2/C2. The difference is that in this game, one of the equilibria (viz., R1/C1) is Pareto-preferred to the other. In the R2/C2 equilibrium, Row and Column receive a payoff of 1. But in the R1/C1 equilibrium, both players would receive a larger payoff of 2. Given that Row and Column are in the equilibrium where both players choose strategy 2, there is an alternative state in which at least one player would be better off without making any other player worse off—namely, the R1/C1 equilibrium. Thus, equilibrium R1/C1 is Pareto-preferred to equilibrium R2/C2, and R2/C2 is Pareto-sub-optimal with respect to R1/C1.
Nevertheless, R2/C2 is still an equilibrium. For if Column chooses C2, Row’s best reply is to choose R2. And if Row chooses R2, Column’s best reply is to choose C2. The R2/C2 equilibrium is self-enforcing, because no player has an incentive to deviate from it. Thus, even with the knowledge that R1/C1 yields better payoffs, the players might be “trapped” in the Pareto-sub-optimal equilibrium. Assuming a social norm is responsible for making Row and Column rationally coordinate on R2/C2, the norm would be unpopular. Both players would prefer to coordinate on R1/C1. However, it is not rational for either to switch to strategy 1 absent some guarantee that the other player will do the same.

In Mackie’s view, the FGC norm is unpopular because most people who practice and endorse it, both men and women, would prefer that it be abolished if they were informed of its consequences. He suggests that for all daughters to be uncut is the best state of affairs for (almost) everyone in an intermarrying community. In this state, there would be no stigma against being uncut. Each uncut daughter would retain marriage-ability and family status, while avoiding “harm to health and violation of human rights” (Mackie, 1996, p. 1008; Mackie & LeJeune, 2008, p. 12). It is also best for men to marry uncut women. The common belief that FGC is necessary to assure fidelity is “self-enforcing”; the very fact that genital cutting is said to be needed to curb female sexual desire creates the perception that women are licentious, and thus unable to resist sexual temptation unless they are cut. In fact, the monogamous male likely has small fidelity-assurance costs. Furthermore, if men desire to successfully raise their own children, as
Mackie’s account assumes, they have reason to prefer that their wife’s fertility, productivity, and longevity not be damaged by genital cutting. Therefore, the unpopular state of affairs is the one where all families in the intermarrying community make their daughters undergo genital cutting.

Despite its unpopularity, the FGC norm is an equilibrium. It may be rational to cut one’s own daughter if enough other people cut theirs. (Analogously, it would be rational for Row to choose R2 if Column chooses C2, despite the unpopularity of the R2/C2 equilibrium.) All actors prefer marriage to non-marriage, and having some convention(s) for signaling spousal quality is preferable to having none. Given that the rest of an intermarrying community sticks to the FGC convention, a family that unilaterally abandons FGC is worse off. For then its daughters would lose marriage-ability and status (along with the whole family). It is only the collective abandonment by most or all families within an intermarrying community that would make it rational for a single family not to cut their daughters. The problem, as Mackie sees it, is to explain how instrumentally rational individuals might cooperate to bring about a collective shift from the cutting equilibrium to a non-cutting equilibrium.

Mackie offers the following account of how an inter-marrying community can escape an unpopular FGC norm (Mackie, 1996, pp. 1011 – 1014). The account assumes that all actors are instrumentally rational. The relevant actors are individuals—men or

21 Originally, in his 1996 paper, Mackie applied this analysis to Chinese footbinding, a thousand-year old practice which ended in only a generation. Mackie sees significant parallels between footbinding and FGC, and elsewhere he uses the same analysis to model the abandonment of FGC (see UNICEF, 2007).
women, boys or girls. We start with a state of affairs in which an FGC norm is being followed by an intermarrying community composed of many individuals. The FGC norm is an equilibrium, since it’s rational to follow the norm given that enough others follow it. What constitutes “following” the FGC norm depends on one’s place in a family: if one has a daughter, following the norm will be to have her cut. If one is a daughter, following the FGC norm is to get cut. If the individual is seeking out a wife for himself or for his/her son, to follow the FGC norm is to only select cut women as prospective wives.

Next we suppose that there is a minimum number of people, call it the critical mass, who would be better off by pledging not to cut their daughters on the condition that other members of the critical mass do the same (Mackie, 1996, p. 1012; Mackie & Lejeune, 2008, p. 14). The critical mass is a subset of the intermarrying community, and it need not be a majority. For those in the critical mass, the advantages of pledging against FGC outweigh the costs of a smaller mate-selection pool (Mackie, 1996, p. 1012). This would be the case if the members of the critical mass are motivationally atypical in terms of how well FGC serves their desires. Compared to the rest of the community, these atypical individuals may receive less utility from following the FGC norm. They may, for instance, be more sensitive to the adverse health consequences of FGC. They may also receive relatively more utility from following a norm against cutting. Perhaps they would be more satisfied by seeing their daughters live without the pain, illness, and infertility that so often result from FGC. But even for individuals who are motivationally atypical in this way, there has to be a minimum number of similarly-motivated individuals who are
willing to make the anti-FGC pledge as well. When the number of anti-FGC pledges has reached this minimum number, there will be a large enough mate-selection pool to make it more advantageous for atypical individuals to follow a non-cutting norm. Call this minimum number $k$. The critical mass is composed of at least $k$ individuals.

Once a critical mass of $k$ or more individuals has established a pledge association, it will be in the interest of the association’s members to get more people to make a pledge against FGC. This is because each new recruit increases the mate-selection pool for the members of the pledge association. Members of the pledge association may be tempted to renege on the pledge. But the pledge itself can be enforced by a social norm. If enough members of the pledge association are motivated to punish defectors, and the costs of these punishments exceed the benefits of breaking the pledge, then it will be rational for each association member to continue honoring the pledge.

The first condition for a shift from a cutting norm to a non-cutting norm is that a critical number of at least $k$ people pledge not to practice genital cutting. The second condition is that the pledge association succeeds in recruiting more members, and controls the temptation to renege through sanctioning. The third, decisive condition is that the pledge association expands to a critical number of individuals called the tipping point, labeled $t$. When $t$ individuals have pledged not to practice FGC, the benefit of making the anti-FGC pledge equals the cost of ceasing to practice FGC even for typical members of the community. This might occur when the pledge association offers a large enough mate-selection pool to compete with the pool of those who continue to practice
FGC. It could also occur when the pledge association, through its recruitment campaign, has changed enough people’s perceptions of the relative costs and benefits of FGC, so that the value of abandoning FGC is perceived to be greater relative to the value of maintaining the practice. Whatever the case may be, when the pledge association has attracted \( t + 1 \) individuals or more, the benefit of committing to the anti-FGC pledge will exceed the cost of abandoning the FGC norm, even for a typically motivated individual. In that case, it will be rational for motivationally typical individuals in the community to pledge against FGC, and to punish those who continue practicing it. In this way, an intermarrying community can achieve a shift from the unpopular FGC norm to a more popular norm against genital cutting. Of course, it’s unrealistic to conclude that the anti-FGC norm would be unanimously preferred to the cutting norm. There might be other atypically motivated individuals—the counterparts of the critical mass—who are unusually satisfied by the practice of FGC. For them, it will never be rational to pledge against FGC. But we can assume that, like the individuals in the critical mass, they are a minority population. The anti-FGC norm may not be unanimously preferred or Pareto-preferred to the FGC norm, but it can at least be more popular.

Mackie assumes that the cutting norm is less popular than the non-cutting norm. However, the norm’s unpopularity is not essential to the account of how it might be escaped. In principle, the logic of Mackie’s account can also explain a shift away from a popular norm to a relatively unpopular one. Remember: norms are systems of expectations. The primary motivation for following norms and conventions is that there is
a payoff for acting as other people expect you to. Mackie’s convention account shows how a new system of expectations can emerge to challenge an existing one. There are both costs and benefits associated with following the new norm versus the old one. But if the new system can recruit an increasing number of followers, it will become increasingly costly for the undecided to keep following the old one. Mackie’s theory implies only that it takes fewer recruits to shift from an unpopular norm to a more popular one than vice versa. For instance: in Mackie’s model, the opportunity cost of joining the association of people who pledge against FGC is a reduced selection of mates. By ceasing to practice FGC, one foregoes the opportunity to inter-marry with people who do practice it, and where FGC is common these people will be a majority of the population. Additional costs of abandoning FGC that are not explicitly included in Mackie’s model include contempt and ridicule meted out by the FGC practitioners. For an individual to rationally abandon FGC, these costs must be balanced by the advantages of following a non-cutting norm. But if individuals have a preference for the state of affairs in which the non-cutting norm is generally followed (say, because of the health advantages), then this weighs in favor of pledging to abandon FGC. Consequently, the pledge association against FGC would not need to offer one prospective mate for every prospective mate foregone by joining the association. And in that case, the pledge association could attract new recruits without having to be as large as the group that continues to practice FGC. The more a typical individual prefers following a non-cutting norm, relative to a cutting norm, the fewer recruits that are necessary to reach the critical mass, and ultimately, the tipping point. For
this reason, the unpopularity of a norm can be a source of instability. But again, a population could in principle shift away from any norm, popular or unpopular, as long as enough members of the population participate in the shift.

Mackie (1996) originally proposed this theory of convention shift as an explanation for the sudden demise of footbinding in China. But in his 1996 article, he also predicted that the same process could bring about a quick end to FGC (Mackie, 1996, pp. 1014 – 1016). This prediction has apparently been borne out by successful anti-FGC campaigns which used the pledge technique. For example, in the late 1990s a nongovernmental organization (NGO) by the name of Tostan implemented a “basic education” program in several villages throughout Senegal. The eighteen-month program was oriented toward women, though men were not excluded. The program included learning modules on a diverse array of topics, including financial management, prevention of child mortality, women’s health, and human rights (Mackie, 2000, p. 259; Mackie & LeJeune, 2008, p. 32). In the Senegalese village of Malicounda Bambara, a Tostan facilitator broached the topic of FGC to a group of thirty-nine program participants. At first, the women refused to speak on the matter. But after a few days they started to respond to the facilitator’s questions and compare experiences with FGC. Eventually the participants collectively discovered a link between FGC and negative health consequences. For example, one woman from a nearby village who had once been a cutter revealed that she had resigned her position because her own daughter was almost killed by the procedure. Apparently most of the women had not correlated gynecological
and obstetric problems with FGC, instead believing such complications to be either isolated incidents or “normal.” In view of these breakthroughs, the women then decided to make it a village project to end FGC. On July 31, 1997, Malicounda Bambara declared publicly that it would abandon FGC, and urged other villages to follow their example (Mackie, 2000, p. 256). According to Mackie, this is the first unequivocal collective abandonment of FGC on record.

At first, the Malicoundan declaration was very controversial—so much so that people from other villages sent hostile messages to Malicounda. The Malicoundan women were hurt by the negative reactions, but they defended their position and even traveled to the villages of Nguerigne Bambara and Keur Simbara to discuss their views with other women participating in Tostan’s basic education program. The example of Malicounda proved to have some effect in leading other villages to abandon FGC. On February 14 and 15, 1998, delegations from thirteen villages representing 8,000 people—including Nguerigne Bambara and Keur Simbara—came together and issued the “Diabougou Declaration.” The Diabougou Declaration expressed a commitment to “end the practice of female ‘circumcision’ in our community” and to “spread our knowledge and the spirit of our decision to our respective villages and to other communities still practicing female ‘circumcision.’…” (Mackie, 2000, p. 257). Then, on June 1-12, 1998, representatives of eighteen villages in the Kolda region of southeastern Senegal, led by the village of Medina Cherif, which had completed Tostan’s basic education program, issued the “Medina Cherif Declaration.” The Declaration states that “we hereby join the
great movement to end female ‘circumcision’ which began in Malicounda Bambara” (Mackie, 2000, p. 258). News of these first declarations spread throughout Senegal and the world. Dozens more villages went on to make similar declarations, and on January 13, 1999, the government of Senegal enacted legislation prohibiting FGC (the effects of which have proven to be mixed) (Mackie, 2000, p. 260).

This story follows the pattern of Mackie’s theory of escaping unpopular conventions. The NGO Tostan was able to establish a critical mass of individuals in Malicounda willing to commit to a pledge against FGC. Mackie suggests that the pledge in Malicounda was enforced by informal sanctioning: “public opinion continues to resolutely oppose FGC, and villagers say that deviators will be identified and punished” (Mackie, 2000, p. 257). The Malicoundan women endured the sanctioning of others, and contributed to successful efforts to recruit more members to the pledge association from other villages. It’s difficult to say whether Tostan’s campaign is nearing a tipping point in favor of the universal abandonment of FGC in and around Senegal. However, Mackie asserts that the work of Tostan and other organizations has generated “realistic hopes that [FGC] can be eliminated within one generation” (Mackie & LeJeune, 2008, p. 1).

Mackie’s theory of norm-shift informs an answer to the opening question of this section: Can it be rational for members of a community to violate a moral norm that is generally followed? I suggest the answer is “yes.” Imagine a young woman, Fauziya, in
an intermarrying community that practices FGC. Fauziya detests the practice. She went through the procedure herself, and has experienced first-hand the health complications and pain associated with FGC. Also, having made contact with NGOs working to eradicate FGC, Fauziya is convinced that genital cutting is a violation of women’s human rights. She vows to end the practice of FGC in her village once and for all. However, Fauziya knows that her personal opposition to FGC is atypical. Most people in her community rationally follow a norm which requires genital cutting. If Fauziya is to achieve her goal, she has to foment changes in the system of expectations that would ultimately make it rational for (nearly) everyone to follow a non-cutting norm. First, she must find a critical mass of villagers for whom it would already be rational to make conditional pledges to abandon FGC (on the condition that enough others do likewise). Furthermore, it must be rational for members of Fauziya’s association of dissidents to persuade others in the community, who are less hostile to genital cutting, that the practice should be abolished. The members of the association should also be willing to deter defectors through punishment. Finally, Fauziya knows that the universal abandonment of FGC will require that the number of people in her pledge association surpass a tipping point beyond which it will be rational for nearly everyone else in her community to abandon FGC. If Fauziya stands a reasonable chance of success in bringing about these

22 “Fauziya” is the first name of a young woman, Fauziya Kassindja, who was granted political asylum in the United States after fleeing her home country of Togo to escape genital cutting (Nussbaum, 1999, in Feinberg & Shafer-Landau, 2008).
events, it may be within her power to change her social environment in a way that makes it rational for her fellow villagers to shift to a non-cutting norm.

Instrumental rationality does not necessarily demand obedience to whatever moral norms prevail in one’s community. Of course, there are atypically motivated individuals, like Fauziya, who would prefer not to follow a given norm, no matter what. But even a typical individual who follows a status quo norm can rationally convert to a new norm, provided that sufficiently many others do the same. And if the status quo norm is unpopular relative to an alternative norm, it will be easier to persuade typical individuals in the community to abandon the status quo in favor of the alternative.

5.6 Persuasive communication, “fundamental moral norms,” and basic human needs

5.6.1 Internalization by persuasion

Although I think Mackie’s theory of norm-shift is correct in its broad contours, it still seems to underestimate the considerable power that norms have to motivate compliance. In particular, it underplays the compliance and punitive motivations that conspire to ensure conformity. The main problem is that the theory of norm-shift described above is really a story of convention-shift. It’s already been mentioned that conventions can become norms over time. But when this happens, whatever motivation there is to follow the convention will be augmented by motivations specific to the psychology of norms. These include a conditional preference to conform for its own sake (so long as enough others do likewise), and a preference to avoid punishment for
violating norms. For example: if FGC were a mere convention, then the sole cost of refusing to practice FGC would be a reduction of prospective mates for oneself or one’s son/daughter. FGC is not just a convention, however. In some communities it is a full-fledged moral norm. It is associated with a system of normative expectations rooted in shared moral and aesthetic judgments. When a girl or woman is not cut, she has to endure the stigmatization of her peers. They may consider her ugly, immodest, unclean, and licentious. She has to live with the thought that she diminishes the honor and standing of her family, or that of her husband’s family. And if she has internalized these normative expectations, she may feel tremendous guilt and shame for not undergoing the procedure. It seems these factors would make norm-shift much more difficult to achieve. Even if a person’s desires would be better served by a different norm, a unilateral shift away from the prevailing norm can entail very high costs that would not be compensated merely by an alternative mate-selection pool. These considerations raise doubts about whether it can ever be rational to dissent from the prevailing moral norms, even if they are unpopular. In what follows, however, I suggest that some norms may be so detrimental to one’s basic human needs that one’s best instrumental reasons will favor being a moral dissident.

Norm-shifts do happen, but there must be additional mechanisms which explain what attracts individuals to the emerging norms and away from the prevailing norms. In more recent work, Mackie refines his theory to better account for the process by which a person internalizes a new norm (Mackie & Lejeune, 2008). At every stage in the course of norm-shift, individuals within an intermarrying community engage one another in
persuasive communication about which among a set of alternative norms the community ought to follow (Mackie & Lejeune, 2008, p. 15). The effect of such communication is to increase the individual’s preference for a state of affairs where a new norm is followed relative to the state where the status quo norm continues to prevail.

Persuasive communication occurs under several conditions. First, individuals must be informed about alternative norms and their consequences. Second, if a norm-shift is to occur, the discussion of alternatives should increase the individual’s preference for the new norm relative to the status quo norm. When this happens, a smaller portion of the community is needed to reach the tipping point which makes it rational for the rest of the community to follow the new norm. A preference for following the new norm is strengthened in either of two ways: the utility one receives from the status quo norm (e.g., FGC) decreases as a result of information about its relative disadvantages (e.g., adverse health effects), or the utility of the new (non-cutting) norm is increased in light of information about its relative advantages (e.g., health, human rights concerns) (cf. Mackie, 1996, p. 1011).

Persuasive communication plays a major role in fixing for an individual what aspects of following a given norm constitute an “advantage” or “disadvantage.” Advantages and disadvantages may be framed in terms of norms that the individual has already internalized, or in terms of the alternative norms under consideration. However, in the next section I argue that the norms which are most likely to be accepted as “advantageous” are those which best promote what I call one’s basic human needs. These
are desires which are common among human beings, and which are not contingent on any particular set of norms that an individual has internalized.

The first condition for persuasive communication is being informed about the alternatives and their consequences. This was an important barrier to the abandonment of FGC in many African communities. In areas where FGC is heavily practiced, being uncut is not even considered an alternative. For instance, interviews with Malian women show that many had never even thought of being uncut as an option (Mackie & Lejeune, 2008, p. 19). On the other hand, in areas with lower levels of FGC prevalence, where there is more exposure to non-practicing groups, there is declining support for FGC among younger age groups. No such trend is found in countries where FGC is highly prevalent (Mackie & Lejeune, 2008, p. 20). The experience of NGOs working to end FGC also confirms the importance of being informed about alternatives. The NGO Tostan found that the first collective abandonment of FGC in Malicounda, Senegal was the most difficult. But each successive one in the country became easier because, as Mackie observes,

> [t]hose deciding whether they too should abandon the practice can send delegates to successful villages...to verify...that an alternative is possible, that others like them have concluded that the alternative is valuable, that collective abandonment works, that daughters’ reputations remain intact, that the effort brought pride rather than shame to the community, and that no supernatural penalties occurred (Mackie & Lejeune, 2008, p. 23).

Once the alternatives are defined, persuasive communication involves the collective assessment of advantages and disadvantages. The health complications associated with FGC have consistently been perceived as a disadvantage among communities considering
abandonment. The Malicoundans, who achieved the first documented collective abandonment of FGC, cited negative health effects as the primary motivation for ending the practice (Mackie, 2000, p. 258). Often the women in areas where FGC is common do not even associate these hazards with genital cutting. Instead, they view them as normal—“the common lot of women”—or as the result of other causes. In Medina Cherif, Senegal, when girls died, bled, or got infected at cutting, the women there attributed it to evil spirits (Mackie, 2003, p. 147). But upon hearing from a source they considered credible that there was a causal relationship between complications and FGC, a group of local women from Medina Cherif had the following reaction (Mackie, 2003, pp. 147 – 48):

…it took them thirty minutes of discussion to decide that the causal claim was correct. They reviewed local history and suddenly realized that incidents of death, hemorrhaging, and infection were immediately associated with [FGC], and they broke down and wept.

Apparently, being persuaded of the negative health effects of FGC has strengthened a preference for a non-cutting norm.

Another way to strengthen one’s preference for a non-cutting norm is to reflect on its advantages. Whether or not a norm is perceived as “advantageous” is, I argue, determined by the norms that one has internalized. In the case of FGC, it seems that the introduction of human rights discourse into the deliberations of several communities hastened the abandonment of genital cutting. Mackie observes that “[i]n the first observed mass abandonments, in both Egypt and Senegal, public commitments to end [FGC] came only after human rights deliberation was introduced into their basic
education curricula” (Mackie & LeJeune, 2008, p. 32). For example, in Deir El Barsha, Egypt, the non-governmental organization CEOSS\textsuperscript{23} projected FGC as a violation of the rights of women and girls “to live a life free of physical and moral violence” (Ibid.). Deir El Barsha was also one of the earliest communities to achieve collective abandonment of FGC.

Mackie hypothesizes that human rights discourse makes salient what he calls a \textit{fundamental moral norm} which, when judged to be in conflict with FGC, motivates abandonment (Mackie & Lejeune, 2008, pp. 28 – 29). He suggests the fundamental moral norm in question is one which obliges parents “to care for their children, not to do them harm, or when constrained, to choose the lesser harm for them” (Ibid.). Fundamental moral norms are, for Mackie, distinct from “less fundamental social norms derived from them.” The former are said to be commonly found across different cultural groups, and are more enduring over time. By contrast, derivative social norms instantiate fundamental moral norms in particular social circumstances, and so they tend to vary from group to group and change over time (Ibid.). Mackie suggests that the fundamental moral norm against harming one’s own children is implemented in some cultural contexts as a norm that requires genital cutting for daughters. This occurs under circumstances where genital cutting is interpreted as a signal of spousal quality in a marriage-ability convention. In a social environment where being cut is considered a condition of marriage-ability, and getting married is an important determinant of a woman’s economic security, it is

\textsuperscript{23} Coptic Evangelical Organization for Social Services.
preferable for one’s daughter to be married and cut than unmarried and uncut. So, conscientious parents in this environment would fulfill their fundamental obligations by having their daughters cut. This explains why mothers so often justify the practice of FGC by appeal to the responsibilities of a good parent.

Mackie’s contention is that human rights discourse provides an evaluative framework which allows them to better appreciate the hazards of FGC. That said, Mackie doesn’t think human rights are a culturally alien system of values imposed on the African communities by activist organizations bent on proselytizing the world to Western humanism. Rather, he maintains that the peoples who pledge to collectively abandon FGC “are not changing their most basic values, but, in response to credible new information, they are realizing them more coherently and fully” (Mackie & Lejeune, 2008, p. 29). The international human rights framework has been appropriated, not assimilated, by the men and women in abandoning communities as a means of articulating opposition to FGC. At the same time, Mackie insists that human rights principles are consonant with the moral traditions of the abandoning communities.

The fusion of human rights with local moral traditions may be a key factor in bringing about a shift from a cutting norm to a non-cutting norm. This section began with the worry that the costs of violating an FGC norm for the sake of shifting to a non-cutting norm would be prohibitively high, as those who pledge to abandon FGC would be subject to intense sanctions from those who continue with it. This is precisely what happened just after Malicounda publicly declared its abandonment of FGC. But the Malicoundan
women were resolute in the face of criticism. They traveled to other villages, argued successfully for their stance on FGC, and ultimately helped encourage other communities to follow their example. These events suggest that the process of recruiting new members to an anti-cutting pledge association cannot be reduced to the offer of an alternative pool of potential mates, either taken or denied. Instead, it’s a process whereby the recruit internalizes an alternative set of moral norms after discussing them thoroughly with trusted peers. A similar process seems to operate in public goods experiments with communication, where the group decides upon an “appropriate” level of contribution. Having internalized the alternative norms, the new recruit will rationally commit to her (his) pledge out of a moral desire to follow the new set of norms. Furthermore, the recruit will not be so sensitive to the censure of the old guard that she’ll only retreat back to the practice of FGC at the first sign of criticism. On the contrary, she will be prepared to defend her newfound convictions. Whatever the costs involved in abandoning a status quo moral norm, they are balanced out by the benefits of serving newly internalized moral desires to follow the alternative norm.

5.6.2 Basic human needs as Sperberian biases

Mackie never fully explains why human rights discourse in particular would help tip communities towards abandoning FGC. But this is something which needs to be explored further. He suggests that the norm against not harming one’s children is a “fundamental norm,” whereas the FGC norm is merely a “derivative social norm.” Information about the health consequences of FGC and human rights discourse help bring
about a shift away from a cutting norm because it exposes a tension between the fundamental norm and FGC. However, it’s not clear why there has to be a tension in the first place. Even granting that all the villages practicing FGC had “fundamental moral norms” against harming their own children, why wouldn’t some villages simply continue to practice FGC despite knowing of the health hazards, and justify it as a means of making girls and women honorable and beautiful? Why wouldn’t parents calculate that the stigma of being uncut (therefore morally suspect, ugly, dirty, etc.) is a worse fate than the health risks of being cut? In that case, the duty not to harm one’s children would favor FGC. Furthermore, they could reduce the negative health effects of FGC by having it performed by trained medical personnel. Indeed, the medicalization option is becoming increasingly popular in countries where FGC is practiced. The important question, then, is why communities that receive human rights education tend to opt for abandoning FGC, rather than adjusting it. To better understand the mechanisms underlying norm-shift, we need to look at the features of human rights norms that make them contributing factors in the abandonment of FGC. This will shed further light on the conditions under which it would be rational to violate prevailing norms in favor of alternative ones.

To better account for the transformative impact of human rights on FGC communities, we should inquire further into the process of norm acquisition. Recall that Stich and Sripada hypothesize that the psychology of norms is underwritten by two mechanisms: a norm acquisition mechanism and a norm implementation mechanism. The

function of the acquisition mechanism is “to identify norms in the surrounding cultural environment, infer their content, and pass that information along to the implementation component” (Stich & Sripada, 2006, p. 298). Consider a child being brought up in a given cultural environment. His norm acquisition system might be entirely unconstrained, in the sense that it will internalize all and only those norms that are generally followed in his environment. This is an unlikely scenario, however. The fact that collective abandonments of FGC norms occur indicates that adults are capable of internalizing new norms. A plausible hypothesis, which I think is more congruent with Mackie’s and other lines of research, is that the acquisition system is biased in favor of internalizing norms that serve basic human needs. This hypothesis is a version of a theory of norm-acquisition which Stich and Sripada dub the “Sperberian biases” view. Named for the anthropologist Dan Sperber, this view holds that “some sorts of normative rules may be more or less ‘attractive,’ due to the way they interact with one’s preferences, aversions, emotions, and other elements of one’s psychology” (Stich & Sripada, 2006, p. 300). Sperberian biases are features of individuals’ psychologies—preferences, aversions, emotions, etc.—which favor the internalization of certain kinds of norms and disfavor the internationalization of others. These biases constitute a selective force in the cultural evolution of norms. Cultural evolution is a process whereby cultural variants—ideas, skills, beliefs, attitudes, values, and norms—are acquired, adjusted, and transmitted among socially interacting individuals over time (Richerson & Boyd, 2005, p. 59).
Cultural evolution is directed by a number of *selective forces* which increase or decrease the probability that a given cultural variant will continue to exist.

Sripada has independently defended a prominent version of the Sperberian biases account (Sripada, 2008). He argues that innate structures in the human mind cause norms to be perceived as more or less “attractive” to the individual, and as a result more or less likely to be acquired, internalized, transmitted, etc. For instance, Sripada takes note of the *Westermarck mechanism*, which is said to generate an innate aversion to having sex with people one has had close, extended associations with during one’s childhood years (Sripada, 2008, pp. 334 – 338). This mechanism is thought to be responsible for the fact that even unrelated children who grow up together in prolonged intimate associations are extremely unlikely to marry or have sexual relationships with one another. It’s believed that the Westermarck mechanism performs the evolutionarily adaptive function of making humans psychologically averse to incest. Since the people with whom one has close, prolonged associations during childhood years are likely to be genetic kin, the Westermarck mechanism would have been favored by natural selection. Sripada argues that the Westermarck mechanism explains the near-universality of incest prohibitions. That is, almost all human societies have moral norms prohibiting sexual intercourse between members of the nuclear family (Sripada, 2008, p. 323). If a moral

---

25 Shaun Nichols (2004) also advances a sophisticated Sperberian biases account. On his view, norms which proscribe actions that elicit negative affect are more likely to persist over time than norms which are not “affect-backed.” Nichols presents historical evidence to show that norms forbidding actions that provoke disgust and empathy (e.g., norms against spitting in public and cruelty to animals) are likely to become common.
norm permitting incest between members of a nuclear family ever emerged, the Westermarck mechanism would impede this norm from being generally followed within a population.

On the other hand, Sripada thinks the Sperberian bias view is also compatible with significant cross-cultural diversity in moral norms. For instance, the Westermarck mechanism does not absolutely preclude the emergence of a norm permitting incest among nuclear family members. According to census records from Graeco-Roman Egypt, more than 30% of marriages were between siblings in some urban areas. Also, historical texts suggest that the Zoroastrian tribe of ancient Iran condoned both sibling incest and parent-child incest (Prinz, 2009). The innate biases view does not assert that incest-permitting moral norms can never arise. Rather, the theory predicts that if an incest-permitting norm arises, it will probably not be followed by a majority of the relevant population, and even if it is generally followed, it will probably be abandoned in favor of norms that are more congruent with incest aversion. So, although incest-permitting norms can emerge, Sripada’s theory predicts that they will be uncommon and relatively short-lived. Sripada’s Westermarck hypothesis is substantiated by the fact that moral norms forbidding sex with nuclear kin are typical across human societies.

In the remainder of this section, I will sketch and give preliminary evidence for a Sperberian bias account which helps to explain the collective abandonment of FGC norms. This account supports the hypothesis that human rights discourse contributes to abandonment because human rights norms which forbid genital cutting are more
consonant with peoples’ basic human needs than genital cutting norms. Basic human needs, as I understand them, are desires to achieve five general types of ends: (1) positive affect, (2) physical security, (3) autonomy, (4) relatedness, and (5) competence. The basic needs are species-typical in that most human beings are motivated to achieve them.

Psychologists have offered an assortment of theories of basic human needs, with considerable evidence gathered in support of each theory. The various theories differ along various dimensions (for a review, see Pittman & Zeigler, 2007). In particular, they posit different lists of basic needs. The alternatives may be compatible, however. Attempts to compare and integrate the disparate theories are underway (Pittman & Zeigler, 2007). Also, there is considerable overlap in the needs hypothesized under the different theories. All five of the needs I have identified as basic are recognized by at least one major research program on basic needs.

The criteria I shall apply in designating a basic need are relatively spare. First, a putative basic need must satisfy the commonness criterion. A basic need is a type of end which is commonly desired by most human beings. However, basic human needs are not necessarily possessed by all humans. The non-universality of basic needs does not make them any less basic. Basic human needs are important theoretically, because they can serve to explain the variety and complexity of human behavior by appeal to a few

fundamental motivations (Pittman & Zeigler, 2007). Since basic needs are invariant across human societies, most theorists take them to be innate features of evolved human psychology. They are not to be explained by a process of enculturation or of internalizing the prevailing norms in one’s immediate social environment. Instead, many prominent theorists give evolutionary rationales for the innate needs they postulate. On the innatist view, basic needs are motivational mechanisms forged by natural selection as solutions to recurrent adaptive problems that humans faced in ancestral environments (cf. Deci & Ryan, 2000; Corning, 2000; Pyszczynski, 2000). However, I will not concern myself with arguing that basic needs are innate adaptations. Indeed, I even grant the possibility that basic human needs are not products of our innate psychological architecture. They may instead be rooted in convergent *cultural* evolution. However the basic needs originated, my central claim is that they bias the acquisition, internalization, and transmission of norms.

The second criterion which distinguishes basic needs from less fundamental desires is that they should contribute significantly to an agent’s overall ability to satisfy his or her desires. Call this the *centrality* criterion. Again, basic needs are supposed to explain a wide variety of behaviors by appeal to a relatively small number of motivations. Thus, basic needs are expected to generate a broad assortment of behaviors, in a wide range of settings (Baumeister & Leary, 1995, p. 498). We might be able to explain why a

---

Moreover, some researchers have expressed hope that the theory of basic human needs can make possible a division of labor between *adaptationist* and *culturalist* explanations of human behavior (Sripada, 2008b; Deci & Ryan, 2000, p. 230, pp. 255 – 256).
philosopher works feverishly on multiple revisions of a manuscript by a desire to publish something important. We might also be able to explain why that same philosopher practices the piano intently every day by a desire to hone his musical skills. A basic human need could provide a unified explanation for both these desires. This philosopher, like many people, has a need for competence—he wants to successfully carry out challenging projects which he values. A theory of basic needs identifies the unifying features which explain why desires and other conative states have motivational power at all. Thus, if an agent is unable to meet one of his basic needs, we should expect his overall ability to fulfill his desires to be dramatically impaired. A person who cannot fulfill one of his basic needs is one who cannot achieve a great many ends that she desires.

The third criterion which distinguishes basic human needs is that they should bias the cultural evolution of norms. If so, then they should play a role in explaining large-scale norm-shifts. We can call this the selection criterion, since basic needs are said to be selective forces which influence the probability that a given norm will emerge and/or persist. Again, a bias does not guarantee that any particular norm will or will not arise. It is rather one force among many others which determine the probability that a given norm will become prevalent in a population. Norms that permit or enable people to satisfy their basic needs are more likely to emerge and persist than norms that do not.

Basic human needs are proposed to be the ultimate grounds of human motivation. If the theory of basic needs is true, then the desires of every statistically normal human
being can be explained as arising from at least one of the basic needs. For example: both the philosopher’s desires to publish something important and to practice the piano might be explained as motivations to meet the basic human need for competence. If basic needs are fundamental motivations in this way, then they cannot be chosen by agents. (Otherwise, if they were chosen, then they would not be fundamental because some other motivation would be needed to explain the choice of a basic need.) Accordingly, I take the basic human needs to be unchosen ends along the same lines that David Schmidtz construes the ends of survival and having something to live for (see chapter 4, section 4.4.3). Indeed, I think the basic human needs subsume Schmidtz’s account of unchosen ends. The end of survival can be considered an element of the basic need for physical security, and it seems to me the end of having something to live for is encompassed by the basic need for competence. Given that the basic human needs are unchosen ends, they play the special role attributed to unchosen ends in the account of reflective instrumentalism presented in chapter 4 (section 4.4). That is, they offer “a relatively stable frame of reference” by which one can evaluate all his or her desired ends from the perspective of instrumental rationality (cf. Schmidtz, 2001, p. 245). If, as the centrality criterion stipulates, the overall extent to which a person can expect to satisfy his desires depends significantly on his ability to meet his basic human needs, then an instrumentally rational agent should act in a manner which satisfies his basic needs.

I draw on multiple sources of evidence to show that the basic needs I’ve postulated meet the commonness, centrality, and selection criteria. To show that these
needs are common and central, I discuss experiments and survey research conducted with cross-national samples. To show that the needs meet the selection criterion, I discuss the work of political scientists Ronald Inglehart, Christian Welzel, and their collaborators.

As I present this evidence for the basic human needs, the reader will notice a frequent reliance on a particular indicator known as subjective well-being (SWB) (Eid & Larsen, 2008). This indicator is widely employed by positive psychologists, and its use is expanding into other fields. SWB decomposes into several measures: life satisfaction (i.e. global judgments of satisfaction with one’s own life), satisfaction with life domains (e.g., work satisfaction), positive affect (experiencing many pleasant emotions and moods), and low levels of negative affect (experiencing few unpleasant emotions and moods) (Diener, 2000, p. 34). To measure life satisfaction, Psychologist Ed Diener and his collaborators developed a survey instrument called the “Satisfaction with Life Scale” (SWLS) (Diener, 2009, p. 194). The SWLS asks respondents to rate, on a scale from 1 (“strongly disagree”) to 7 (“strongly agree”), their agreement or disagreement with a series of statements including: “In most ways my life is close to my ideal,” “So far I have gotten the important things I want in life,” and “I am satisfied with my life.” Questions can also be worded to ask about satisfaction with specific life domains, such as work, family life, or education. Survey instruments designed to gauge “affect balance” usually ask the respondent how often they feel positive and negative emotions in a given period of time. Sometimes the respondent is provided with a list of the positive or negative emotions to be assessed. It was found that positive affect and negative affect are uncorrelated,
suggesting that how much of one kind of affect a person experiences has no bearing on how much of the other type he or she feels. In addition, the presence of positive affect and the absence of negative affect are both correlated with life satisfaction, although sometimes they can be at variance (Larsen & Eid, 2008, p. 3). My preference is to use life satisfaction measures as indicators of desire-satisfaction. But since affect is well-correlated with life satisfaction, I will use measures of affect balance and SWB as second-best indicators of desire-satisfaction. SWB typically takes an average of both life satisfaction and affect balance measures.

I consider the satisfaction measures (viz. global life satisfaction and satisfaction in particular domains) to be reliable indicators of the degree to which a person’s desires are satisfied. A natural way to tell whether or not people are “satisfied” with their lives (or some aspect of it) is to ask them. The SWLS questions make the respondent the primary authority in assessing his or her satisfaction level. Moreover, the theoretical construct of “life satisfaction” seems to be cross-culturally valid. According to Tov and Diener, confirmatory factor analyses suggest that “the SWLS measures a single construct, and that the concept of ‘life satisfaction’ may be similarly understood across a wide range of cultures” (Tov & Diener, 2007, p. 697). An additional attraction of the life satisfaction indicator is that it measures something which people judge to be important. College students in 41 nations rated life satisfaction a mean importance of 6.21 on a scale ranging

---

28 For example, consider a “starving artist” who experiences much emotional hardship but is still satisfied to live his life in a way that he values.
from 1 (of no importance whatsoever) to 7 (extremely important) (Diener, Sapyta, & Suh, 1998). Comparing the importance ratings attributed to life satisfaction by American, South Korean, and Chinese college students, Diener and colleagues (1995) found that all three groups rated it as being at least somewhat important (>5.5). However, Chinese students rated life satisfaction as being significantly less important than the South Korean and American samples (Diener et al., 1995, p. 24).

According to the centrality criterion, the basic human needs must contribute greatly to a typical person’s overall ability to satisfy his or her desires. If life satisfaction is a reliable indicator of desire-satisfaction, then the basic human needs should be identifiable as types of goals on which the life satisfaction of individuals strongly depend. Researchers have noted particularly strong correlations between life satisfaction and five goal-types which I enumerate as positive affect, physical security, competence, relatedness, and autonomy. Therefore, I offer these types of ends as candidate basic human needs. I present evidence to suggest that each candidate need satisfies at least the commonness criterion and the centrality criterion. In addition, I review direct evidence that the need for autonomy satisfies the selection criterion, and indirect evidence that the other four needs satisfy this criterion as well.

Positive affect refers to the experience of pleasant or positive emotions. Negative affect is the experience of negative emotions. One study by M.L. Diener and colleagues suggests that across seven regions in the world, a core group of emotions are consistently classed into either positive or negative clusters (M.L. Diener et al., 2004; Tov & Diener,
Positive emotions include *pleasant, cheerful,* and *happy*, whereas negative emotions include *unpleasant, sad,* and *angry*. These findings suggest that people throughout the world classify emotional experience in terms of a positive/negative dimension. However, there is some cultural variation in which emotions are classified as being positive or negative. For example, M.L. Diener et al. (2004) found that *pride* clustered with positive emotions in Latin America, Western Europe, and East Asia, but with negative emotions in Africa, Southeast Asia, and West Asia. Despite this variation, one regularity stands out: there is a strong positive relationship between positive emotions and life satisfaction, and a strong negative relationship between negative emotions and life satisfaction. Kuppens, Realo, and Diener (2008) observed as much in a data set composed of over 8,500 people from 46 countries. On the other hand, Kuppens et al. also found significant cultural mediation of the way emotional experience contributed to or detracted from life satisfaction. In so-called collectivist societies, negative emotions are less dampening to life satisfaction than in so-called individualist societies. *Individualistic* societies are those where the individual is typically seen as the most important social unit. In such societies, the individual is thought to be unique and free to make personal choices that conflict with the desires of others.

---

29 Tov and Diener, however, suggest that the methodology employed by this study may mislead. The authors weighted *frequently experienced* emotions as positive ones. Thus, *worry* and *stress* clustered with the positive emotions in Western Europe and East Asia, primarily because they were frequently experienced in those places (Tov & Diener, 2007, p. 696).

30 In their emotion surveys, Kuppens, Realo, and Diener (2008) listed as positive emotions *pleasant, happy, cheerful, pride, gratitude,* and *love,* and they listed as negative emotions *sad, anger, unpleasant, guilt, shame, worry, stress,* and *jealousy.*
Collectivistic societies are those where people think of the group as the most important social unit. Individuals are expected to carry out strong duties to others and work to promote group harmony, even if it means sacrificing their own desires (Diener, 2008, p. 133). Countries that score highly in individualism include the Western European nations, the United States, Canada, Australia, and New Zealand. Countries that are highly collectivist include the East Asian and Southeast Asian nations (China, South Korea, Singapore, Thailand), and some Latin American and African countries (Chile, Nigeria) (Kuppens, Realo, & Diener, 2008, p. 70). Kuppens et al. suggest that negative emotions are less related to life satisfaction in collectivist countries because, for people in these countries, satisfaction may not only require personal feelings of happiness but also feelings of self-criticism or failure in order to attain more highly valued ideals such as interpersonal harmony. Nevertheless, even in collectivist societies, positive emotions contribute significantly to life satisfaction. Accordingly, an instrumentally rational agent should strive to maintain a high level of positive affect throughout his or her life.

Physical security refers to universal human desires for sufficient levels of material wealth and safety from threats. Regarding wealth, income contributes to life satisfaction across societies (Oishi et al., 1999, p. 981). The mean life satisfaction of a nation is positively correlated with its gross national product per capita. However, the correlation between individuals’ income and life satisfaction is higher in poorer nations than in rich nations. Apparently, people in poorer nations base their life satisfaction judgments more heavily on financial considerations than people in wealthier nations. This suggests that an
extra unit of income contributes more to the life satisfaction of a person who has less of it.\[^{31}\] Furthermore, surveys of homeless people, slum dwellers, and sex workers living in various degrees of poverty in Calcutta, India show these respondents to be slightly negative to slightly positive in their life satisfaction (Diener & Diener, 2001). People who live in poverty have a lower sense of satisfaction than wealthier people, though on average they are not extremely dissatisfied.

Safety is also associated with life satisfaction. Sheldon et al. asked samples of US and South Korean college students to think about the most satisfying and/or dissatisfying event they experienced recently (either in the past month or week), and then rate the degree to which they felt each of 20 emotions during the event. It was found that, in both samples, the subjects associated a pre-defined sense of “security” with the recent satisfying events and the positive affect experienced during those events. They also associated a lack of security with dissatisfying events and negative affect (Sheldon et al., 2001, p. 334 – 336). Sheldon and colleagues defined the sense of “security” as having a structured and predictable life, having a comfortable set of routines, and being safe from threats and uncertainties (Sheldon et al., 2001, p. 328). The evidence for including safety as a component of a basic need for physical security confirms Schmidtz’s conjecture that survival is an unchosen end desired by nearly all human beings (cf. chapter 4, section 4.4.3). If most human beings are motivated to attain safety from threats, then they will be

\[^{31}\] Economist Richard Layard takes these and other findings to be persuasive evidence of the diminishing marginal utility of income (see Layard, 2006).
more likely to survive. Indeed, it’s not implausible to suggest that the desire to pursue safety can be explained by a more fundamental desire to survive.

It is tempting to include a desire for health as an element of the need for physical security. However, it’s inconclusive whether counting health as a part of a basic human need is warranted by the SWB research. In one review, Diener et al. (1999) note that several studies failed to find any effect of “objective” health measures (e.g., health ratings by physicians) on an individual’s life satisfaction rating. Instead, life satisfaction is better predicted by a person’s subjective perception of his or her own health than by the “objective” measures. This possibly explains the finding that the global life satisfaction of severely ill patients (e.g., cancer patients) was only slightly different from non-patients. Also, people in poor health tend to downplay the importance of their health in assessing their life satisfaction. On the other hand, people with symptoms of multiple chronic morbidity conditions (e.g. pulmonary disease, congestive heart failure) showed declining SWB through a one-year period after being discharged from the hospital. And although most people with multiple disabling conditions reported themselves to be satisfied with their lives, they were significantly less satisfied on average than a non-disabled comparison group (Diener et al., 1999, p. 287).

Competence refers to desires to be successful and capable in carrying out tasks, as well as desires for effective interaction with one’s environment (Sheldon et al., 2001, p. 328; Deci & Ryan, 2000, p. 252). In the survey by Sheldon et al. (2001), the sense of competence was strongly associated with satisfying events and positive affect. Also, the
absence of a feeling of competence was very strongly related to unsatisfying events and negative affect. Because competence contributed so powerfully to the experiences of satisfaction and affect, Sheldon et al. conclude that competence is a basic human need (Sheldon et al., 2001, p. 337).

The hypothesis that competence is a basic human need subsumes Schmidtz’s hypothesis that nearly all human beings desire the maieutic end of having something to live for. Schmidtz is likely correct that most people do want something to live for. This is something we want, I suggest, because we want to be competent. And granting that the desire to have something to live for is grounded in the need for competence, we can provide a deeper reply to an important objection that weighed heavily on the reflective instrumentalist view presented in chapter 4. At the end of chapter 4, I considered the worry that a person might satisfy the end of having something to live for by choosing to live for an addiction. This poses a potential difficulty for instrumentalism, since intuitively it may seem that no rational person would choose an addiction. But if people have a need for competence, reflective instrumentalism can accommodate this intuition. The answer to the addiction objection proposed in chapter 4 (section 4.4.4) was that an addicted person likely fails to maximize the satisfaction of her desires. It is well-known that addictions tend to impede people from pursuing other ends that they value and desire. Of course, we can at least imagine a person who desires nothing more than to indulge an addiction. However, if competence is a basic human need, then most people will not have such an anemic range of desired ends. If people have a need to feel
accomplished and effective in their environments, then it’s likely that they will have to pursue a sense of accomplishment in a variety of domains. Whether we like it or not, the natural and social environments that we inhabit impose a broad multitude of demands upon us. To preserve our physical security, we have to effectively cope with our surroundings and attain food, shelter, and safety from threats. In society, we’re expected to be diligent students, capable providers for our children, proficient co-workers, and so on. Most of us are in fact sensitive to these various demands, and we could not attain much of a sense of accomplishment without fulfilling them to some adequate degree. Therefore, I would contend that few people would adequately meet their desired end of having something to live for simply by choosing one end—an addiction, for instance—which they single-mindedly pursue.

*Relatedness* consists of desires to be affiliated with a group, to be caring and altruistic toward others, and to conform to group norms (Deci & Ryan, 2000, p. 253). In the Sheldon et al. study, relatedness was defined primarily in terms of affiliation. It was characterized as “a sense of contact with people who care for me,” as being “connected with other people who are important to me,” and having a “strong sense of intimacy with the people I spent time with” (Sheldon et al., 2001, p. 328). Given this definition of relatedness, subjects in both the American and Korean samples strongly associated it with satisfying events and with positive affect. In fact, the Korean subjects associated relatedness with satisfying events and positive affect more than anything else. Numerous other studies demonstrate that having social relationships is among the most important
determinants of SWB. For instance, the happiest members (top 10%) of one college student sample reported stronger relationships with friends, family, and romantic partners than the unhappiest members (bottom 10%) (Tov & Diener, 2008, pp. 325 – 26). Also, across 42 societies, married people reported more pleasant and less unpleasant affect than the divorced (Tov & Diener, 2007, p. 704).

Evidence for a human propensity to altruism has been accumulated in experiment after experiment over a span of decades (see section 5.4; Frank, 1988; Sober & Wilson, 1998). However, these studies have been conducted primarily in Europe and North America. Cross-cultural research on altruism is relatively thin. Nevertheless, there is some evidence to suggest that other-regarding desires are common among humans in all societies. Henrich et al. (2004) found that individuals across fifteen small-scale societies on average played public goods and ultimatum games in an other-regarding fashion—i.e., contributing generously to public goods and making generous ultimatum game offers.  

More systematically, Tov and Diener analyzed World Values Survey data from 58 different countries and found that average volunteer involvement among a country’s citizens correlates with the country’s mean SWB (Tov & Diener, 2008, p. 333). However, the exact nature of the voluntary involvement was not described in the World Values data.

32 On the other hand, self-regarding behavior was common in some of the groups Henrich visited. The Machiguenga from South America and the Tsimané from East Africa made low ultimatum game offers, and offers were rarely rejected. Interestingly, in daily life these peoples practice little cooperation, sharing, or exchange beyond the family unit.
Finally, a desire to conform to group norms is also cross-culturally ubiquitous. Such a desire would explain why almost all individuals do faithfully conform to the prevailing norms of their group. Psychologist Solomon Asch developed a famous experimental paradigm which elicits the desire to conform. In the typical experiment, participants are faced with a choice between either stating what they saw with their own eyes, or refuting a unanimous group of confederates who express a contradictory opinion. Asch found that most subjects conform at least once by deferring to the group’s opinion. While Asch’s own experiments were conducted primarily in the US, his paradigm was replicated over 100 times in 17 countries. It turns out that subjects living in collectivist countries conform in these experiments even more often than Americans. The proposed explanation is that in collectivist societies, people face stronger expectations and experience stronger motivations to achieve a sense of belongingness with the group (Heine, 2007, p. 727).

The basic need for *autonomy* refers to a desire to act with a sense of free choice and control over one’s own life. The property which distinguishes autonomous from non-autonomous agents is *self-concordance*. A self-concordant agent is one who acts to achieve goals that he or she fully endorses. By contrast, a non-concordant agent does not fully endorse the ends for which she acts. Instead, non-concordant individuals act from the sense of being compelled by external forces. For example, a self-concordant piano student practices the piano because of a genuine desire to master the instrument or cultivate musical skill. She endorses the goals of mastery and musicianship as final ends.
A relatively non-concordant piano student practices the piano because her parents insist on it. She herself, however, has little interest in the instrument (Sheldon et al., 2004, p. 209).

The evidence supporting a universal human desire for autonomy is extensive. Using data from the World Values Survey, Inglehart and Welzel (2005) report a cross-culturally robust correlation between autonomy and life satisfaction. The Values Survey measures life satisfaction on a ten-point scale ranging from 1 (“completely dissatisfied”) to 10 (“completely satisfied”). It also measures the sense of autonomous choice by way of responses to the following question (Inglehart & Welzel, 2005, p. 139):

Some people feel they have completely free choice and control over their lives, while other people feel that what they do has no real effect on what happens to them. Please use this scale where “1” means “none at all” and “10” means “a great deal” to indicate how much freedom of choice and control you feel you have over the way your life turns out.

In each of 191 nationally representative surveys conducted in 73 societies, the sense of autonomous choice and life satisfaction were significantly correlated (Ibid.). The individual-level correlation pooled across all surveys is $r = .40$ ($N = 248,000$). At the aggregate level of nations, the correlation is $r = .78$. Surprisingly, the correlation between life satisfaction and the sense of choice was actually stronger for some collectivist societies than some individualist societies. For instance, the link between life satisfaction and freedom of choice is stronger in China, Indonesia, and Zimbabwe than in the Netherlands, Denmark, Iceland, and Finland (Inglehart & Welzel, 2005, pp. 139 – 141).
It might be wondered whether the basic need for autonomy and the need for relatedness are in tension with each other, insofar as relatedness includes a desire to conform. This need not be the case. People in collectivist societies apparently face stronger expectations to conform. However, Sheldon et al. (2004) found that South Korean and Chinese college students were no less self-concordant than Americans, although a sample of Taiwanese students was less self-concordant than these three groups (Sheldon et al., 2004, pp. 218 – 219). So here we have two samples from countries rated high in collectivism which are just as self-concordant as a sample from a more individualistic country. The lesson to be drawn is that expectations to conform are compatible with the need for autonomy, but only if the individual can internalize the norms of the group to the point that he or she follows them from a sense of free choice rather than a fear of being punished. This presumably occurs when the agent is motivated to comply with norms on the basis of positive or negative internal sanctions—e.g., pride generated by following norms or guilt generated by violating them.

Research on SWB suggests that the five desires I’ve postulated satisfy two of three criteria for being a basic human need: the commonness criterion and the centrality criterion. However, to secure their status as basic human needs, each must meet the selection criterion—that is, they should to contribute to explanations of large-scale social and historical trends. One possibility is that they are a selective force in the cultural

_________________________

33 In all four student samples, Sheldon et al. found that self-concordance was predictive of SWB. This finding replicates that reported by Inglehart and Welzel and discussed in the preceding paragraphs.
evolution of norms. I suggest that if basic human needs operate like Sperberian biases, they would favor the acquisition, emergence, and propagation of norms that permit or enable people to meet their basic needs. On the other hand, they would disfavor the acquisition, emergence, and propagation of norms that prevent people from meeting their basic needs.

To show that each of the basic needs I’ve proposed fulfills the selection criterion is too gargantuan a task to be accomplished here. Fortunately, much of the hard work has already been done when it comes to the human need for autonomy. In a series of books and articles on the process of “democratization,” Ronald Inglehart and Christian Welzel cite “universal human aspirations” for autonomy as a basic engine driving the emergence of democratic political institutions (Inglehart & Welzel, 2005; Inglehart et al., 2008; Welzel & Inglehart, 2009). The work of these authors follows the political culture research tradition in political science. The political culture tradition, whose forerunners include Aristotle and de Tocqueville, affirms that political regimes arise and endure only if they are congruent with the attitudes, beliefs, and desires of the public (Welzel & Inglehart, 2009).

A quick summary of Inglehart and Welzel’s theory of democratization will show how the basic human need for autonomy can explain transitions from authoritarian rule to democracy. The whole process is initiated by the socioeconomic development of a country. Socioeconomic development consists of technological innovations that increase labor productivity, occupational specialization, and rising education and income levels. It
diversifies social interactions and effects major cultural transformations (Inglehart & Welzel, 2005, p. 19). The main effect of socioeconomic development is to increase the material, intellectual, and social resources at the disposal of ordinary individuals (Welzel, 2007, p. 160). As wealth increases, people can invest more time and money in private and civic activities. Furthermore, socioeconomic development is linked with rising levels of education, thus increasing the public’s intellectual skills. Finally, socioeconomic development produces denser, more diversified, and more networked populations, opening up opportunities for people to engage in broad-based interactions such as voluntary associations and, especially, mass political movements. Ultimately, increased material, intellectual, and social resources expands the typical citizen’s repertoire of feasible actions.

Inglehart and Welzel argue that socioeconomic development leads people to adopt *emancipative values*. Emancipative values are ultimately moral convictions that emphasize gender equality, tolerance, freedom of choice, and political participation (Welzel & Inglehart, 2009, p. 130). A person who holds emancipative values is likely to disagree that men make better political leaders, to agree that homosexuality is justified, to prioritize personal choice over obedience to others, and to demand civil protections such as freedom of speech. As Welzel describes them, emancipative values represent “a universal appreciation of human autonomy” (Welzel, 2007, p. 161). On the basis of data from the World Values Survey, Welzel and Inglehart contend that generational shifts
toward generally held emancipative values have occurred in all socio-economically affluent societies (Welzel, 2007, p. 162; Ingelhart & Welzel, 2005, p. 97).³⁴

As emancipative values spread, they tend to instill desires for the civil and political liberties that liberal-democratic institutions provide. These rights and liberties include rights to political participation, civil rights, and equality before the law (Welzel & Inglehart, 2009, p. 133). Insofar as a political regime fails to fulfill these democratic aspirations, it will be viewed as illegitimate by an increasingly skilled, well-resourced, socially interconnected public. Incongruence between mass demands for democracy and present levels of democracy becomes a source of regime instability, as emancipative values motivate the populace to collectively install democratic institutions. Inglehart and Welzel show that the percentage of a country’s population which holds emancipatory values correlates positively with involvement in protest activities (Inglehart & Welzel, 2005, p. 227). Furthermore, analyzing data from the “third wave” of democratization that took place in 1988 – 1998, the two authors demonstrate that when a country has a high proportion of people who embrace emancipative values, given its present level of democracy, it is likely to become more democratic over time. By contrast, when a country has an unusually low proportion of citizens who embrace emancipative values, relative to its present level of democracy, it is likely to become less democratic (Welzel & Inglehart, 2009, p. 139). This suggests that a country’s political institutions are

³⁴ Inglehart and Welzel’s survey data is incredibly wide-ranging. It is collected from 81 societies containing 85 percent of the world’s population during the period from 1981 to 2001.
adjusted to become more congruent with popular demands for democracy. Institutional change is achieved through mass political movements aimed at reducing the discrepancy between freedoms aspired to and freedoms supplied.

Inglehart and Welzel’s whole mechanism for democratic political transitions is set in motion when, as a result of socioeconomic development, emancipatory values become widely accepted in a population. But why should development incline people toward embracing emancipatory values? In response to this question, Inglehart and Welzel advert to a “universal human desire” for choice and self-expression that begins to take priority as people become more economically secure. To substantiate this notion, they appeal to research suggesting that individual autonomy is related to subjective well-being and life satisfaction, while the absence of autonomy is related to lower levels of life satisfaction (Welzel, 2007, p. 157; Inglehart & Welzel, 2005, p. 139). Furthermore, Inglehart and colleagues claim that the desire for autonomous choice is satisfied by democratic institutions. When controlling for other factors (e.g., economic development), democracy predicts the extent to which a country’s population feels a sense of free choice. For instance, according to World Values Survey data collected between 1981 and 2007, people in countries that transitioned from authoritarian rule to democracy during this period reported a rising sense of free choice (Inglehart et al., 2008, p. 270, 280).

Thus, the causal process traced by Inglehart and Welzel’s theory of democratization begins with socioeconomic development and the basic human need for autonomy. In the absence of this basic need, ordinary people would never be motivated to
agitate for democratic institutions that permit them to live autonomous lives. If Inglehart and Welzel’s account is correct, then a basic human need for autonomy clearly meets the selection criterion. Under the appropriate social and economic conditions, a basic human need for autonomy favors the evolution of moral norms that promote emancipative values such as tolerance, equality, and freedom of choice. Inglehart and Welzel endorse an “evolutionary perspective” on democratization (Haerpfer, Bernhagen, Inglehart & Welzel, 2009, p. 382). Democratic systems, they argue, have several “selective advantages” over authoritarian systems that makes the former relatively more stable and more likely to endure. They identify as the most important selective advantage the “human aspiration for freedom,” which makes it “the most demanded system for all people who have acquired the means and ambition to raise their voices” (Ibid.). Democratic systems are seen as legitimate by populations that embrace emancipatory values. Authoritarian regimes that exist at odds with a public committed to emancipatory values have a legitimacy deficit. As a result, they are more vulnerable to destabilizing mass opposition.

Of course, the need for autonomy does not guarantee that democracy will always prevail. An authoritarian regime can repress democratic opposition movements by maintaining control over the means of violence—e.g., the military and the police. China, in particular, is much less democratic than Inglehart and Welzel’s theory would predict, given its high levels of economic growth and moderately high public sympathy for emancipatory values (Inglehart & Welzel, 2005, p. 155). Two decades ago, the Chinese
government thwarted the Democracy Movement of 1989, killing an unknown number of protestors. However, Inglehart and Welzel maintain that China’s government survived the opposition because the Democracy Movement was concentrated among the younger and more educated segments of the still predominantly rural population. They point out that mass opposition has usually failed when led by relatively small, identifiable, and easily isolated groups. But history shows that when mass opposition becomes sufficiently widespread, well-resourced, and sophisticated, even the most repressive authoritarian regimes can be overthrown (Welzel & Inglehart, 2009, pp. 132 – 33).35 The basic human need for autonomy does not make the political development of society a deterministic process invariably culminating with democracy. It is rather one force of social change among many, effective only under certain special conditions and when not overwhelmed by countervailing forces. But it is a force nonetheless, and therefore it satisfies the selection criterion.

Do the other needs I’ve hypothesized meet the selection criterion? I cannot approach that question in full. But the preceding discussion of the need for autonomy gives some reason to expect that the other needs can operate as forces of social change. Like the need for autonomy, the other basic needs correlate significantly with people’s SWB—in particular, their life satisfaction. Inglehart (1990, 1999) found that satisfaction

\[35\] Inglehart and Welzel predict that if socioeconomic development continues in China at its present rate, emancipatory values will permeate the rural segment of the population, the military, and the younger government elites, making it increasingly difficult for the government to resist mass pressures for democracy. Accordingly, Inglehart and Welzel expect China to make a transition to a liberal democracy within the next two decades (Inglehart & Welzel, 2005, p. 191).
with one’s life as a whole is a strong predictor of stable democracy. Moreover, global life satisfaction is also strongly correlated with the level or quality of democracy (as measured by Freedom House scores reflecting the protection of civil and political liberties). The “happier” people are in terms of subjective well-being, the more democratic and more enduring their government is likely to be. To explain these relationships, Inglehart hypothesizes that if people living under democratic institutions are satisfied with their lives, they will tend to support those institutions. Inglehart also found evidence to suggest that low levels of subjective well-being within a country have a de-stabilizing effect on a political system. For instance, SWB sank to unprecedented levels in Russia and Hungary in the years prior to the collapse of the Soviet Union in 1991 (Inglehart, 1999, pp. 108 – 109). Although Inglehart’s arguments are limited to the relationship between SWB and political transitions, he suggests that low levels of SWB reflect a lack of support for the status quo, which can lead to collective efforts to change it. (Inglehart holds in general that transitions to democracy are brought about primarily by the collective actions of mass movements.) If the inability to satisfy basic needs has depressing effects on mass SWB, and people attribute their lack of subjective well-being to the institutions under which they live—political, moral, or otherwise—then we might expect collective efforts to establish institutions that better serve people’s basic needs. In short, the same explanatory role fulfilled by the need for autonomy in the

36 Interestingly, in the years immediately following the Soviet collapse (1995 – 1997), SWB in Russia fell lower still, which raises questions about the stability of democracy in the former Soviet Union (Inglehart, 1999, p. 109).
Inglehart-Welzel theory of democratization might be played by the other basic needs in accounting for other socio-political changes.

5.6.3 Abandonment of FGC: explaining the human rights effect

The foregoing discussion of basic human needs—particularly the need for autonomy—provides a basis for explaining why human rights discourse would help to bring about the collective abandonment of female genital cutting. I suggest that the alternative offered by human rights norms—namely, the freedom to control one’s own body—is inherently attractive because it allows women and girls to better promote their basic needs. It’s perhaps obvious that FGC runs counter to the need for positive affect. Well-known consequences of the procedure include pain, anxiety, and depression, both in the short and long terms. Even if these disadvantages could be avoided by medicalizing FGC, the practice also likely to be at odds with the need for autonomy. Usually the decision to be cut is made for the girl or young woman by members of her family, and often at an age when she is too young to fully appreciate the procedure’s consequences. Furthermore, as Martha Nussbaum notes, FGC is irreversible (Nussbaum, 1999, in Feinberg & Shafer-Landau, 2008). If a woman ever comes to value the capabilities foregone by being cut, there is no turning back. A human rights norm which forbids genital cutting may eliminate the option of receiving the procedure, but this cost may be counterbalanced by advantages that are preserved when one remains uncut (viz., fewer health difficulties, increased sexual satisfaction, more children, etc.). One also suspects that the need for competence might be better served by a non-cutting norm. It’s been
noted that women in societies where FGC is traditionally practiced are expected to bear healthy children for her husband (Mackie & Lejeune, 2008, p. 31). And yet, women who are cut suffer from infertility, and are more likely to miscarry (Mackie, 1996, p. 1003). An uncut woman may thus be more effective in discharging traditional expectations attached to being a wife. Depending on how much a woman internalizes traditional expectations, this may be another attraction of a non-cutting norm. Although it’s not clear that physical health is part of a basic human need for physical security, we still saw that people who suffer from chronic and disabling conditions report lower life satisfaction. It’s well-known that the adverse health consequences of FGC are chronic and disabling. Finally, let us not forget that the basic needs of men might be frustrated by FGC as well. The prospect of daughters and wives living healthier, less painful lives may provoke their empathetic sensibilities. These considerations help to explain why collective pledges to abandon FGC in favor of human rights norms have occurred so quickly and decisively.

5.6.4 Moral compliance and moral dissidence

In section 5.5, our reasons to be moral were identified as reasons to follow moral norms. Conformity with moral norms is instrumentally rational for most human agents because it serves the self-regarding, other-regarding, and moral desires of most agents. But it might be objected that this proposed solution to the “Why be moral?” question holds reasons to be moral hostage to moral norms. This can seem problematic, given the compunctions that many people have about the moral norms which happen to prevail in their own society or in others. In particular, I anticipate that my objectors would have in
mind cases like female genital cutting. They may insist that for a girl or woman to willingly endure this procedure is the antithesis of rationality. And yet my view would seem to imply that it’s almost always going to be irrational for a potential victim of FGC, or those who care for her, to dissent from this practice.

In this section and in section 5.5, I’ve tried to establish that instrumental reasons for action can favor both moral compliance and moral dissidence. Just as moral compliance serves the desires of most individuals, so in many cases can dissidence. To see why, let’s return to the case of Fauziya, who is determined to abolish genital cutting in her community. Can it be rational for Fauziya to agitate for change? That will depend on her chances of success: if the chances are too slim, the effort will futile. But if the chances are reasonably good, and Fauziya is sufficiently resolute, her best instrumental reasons may favor initiating a reform movement. To assess her chances of success, Fauziya needs to reliably estimate the relative preferences of the other villagers for a cutting norm versus a non-cutting norm. This is important, because if the FGC norm is unpopular, Fauziya can expect to have an easier time recruiting enough people to tip the whole village toward abandoning FGC. Assessing other people’s preferences is often difficult, however. But let’s suppose Fauziya has a gut sense that genital cutting is so pernicious that there has to be a considerable number of people who would prefer to end it, as long as they see others doing the same. This intuition is based on Fauziya’s own experience with FGC, and perhaps also the experiences of others. Fauziya has to worry, though, that she lacks an understanding of the typical villager’s attitude toward FGC.
After all, to be successful, it is the *typical* villager that she’ll ultimately need to recruit to her side of the issue. I suggest Fauziya’s intuition about the unpopularity of her village’s FGC norm is likely to be justified. The practice of female genital cutting impedes women and girls from meeting some of their basic human needs. These are needs which Fauziya probably shares with other women and girls in her village. And it’s not unreasonable to suppose the same altruistic inclination driving Fauziya to stop FGC, for the sake of all women in her village, also burns in many of the men. Thus, if her own preferences favor a world in which no woman was cut, then Fauziya can be reasonably confident that the preferences of the others lean in the same direction. Fauziya doesn’t need to master “basic needs theory” to see this; she can feel it in her bones. The theory of basic needs only tells us that Fauziya’s gut intuitions about the other villagers’ sympathy for her cause are likely to be on target. Once she has drawn this conclusion, Fauziya can be assured that her defiance is supported by her best practical reasons.

### 5.7 Conclusion

In this chapter, I’ve attempted to show how moral action can be a rational “best bet.” Most of us have powerful reasons to be moral, because most of us are endowed with a psychology which reconciles our desires with moral norms. We care about how we treat others, and about doing what our community deems to be right or good or just. We care about these things simply in virtue of the fact that we are human. Of course, the most blood-soaked moral catastrophes in history—genocide, apartheid, slavery, etc.—were all collective human feats. The individuals participating in those systems were also
motivated by norms which compelled them to obey prevailing expectations without exception. However, I’ve tried to argue that the connections between desire, norm, and morality are rather strong. Forces in the human psyche have pressed us to adopt norms that promote values of great moral significance—equality, freedom, and tolerance among them. These psychic forces may not be strong enough to guarantee uninterrupted moral progress for the entire human species. But they’ve taken us far enough to ask what reason there is to be moral, amid all the temptations to do otherwise. Indeed, the question wouldn’t make much sense in the first place, unless we were already capable of meeting at least some of morality’s demands. A satisfying answer is in the offing. It is an answer that can be grasped with a fuller understanding of our own human inclinations, and the rules we live by.
References


493


Cummins, D. 1996. Evidence of deontic reasoning in 3- and 4-year old children. *Memory*


504


Terence Cuneo, 312 – 327. Malden, MA: Blackwell Publishing Ltd. (abbreviated “SLC”)


Tov, William and Ed Diener. 2008. The well-being of nations: Linking together trust,


Biography

Andrés Luco was born September 10, 1980 in La Jolla, California. He received Bachelor of Arts degrees in philosophy and modern culture and media from Brown University in 2003, graduating magna cum laude and with honors. The same year, he was elected to the Rhode Island chapter of Phi Beta Kappa. While pursuing graduate studies at Duke University, he held a Duke Endowment Fellowship and was a Kenan Colloquium fellow in the Kenan Institute for Ethics.