Agency and the Foundations of Ethics

Nietzschean Constitutivism

Paul Katsafanas

OXFORD
UNIVERSITY PRESS

Contents

Introduction

Our experience of the world is pervaded by norms. Having promised to meet a friend for dinner, I feel *obligated* to do so. Upon entering a café on the heels of another customer, I think that he *should* have held the door for me. Rousing myself from a relaxing nap, I tell myself that I have *reason* to go to the gym. Watching the evening news, I judge that the criminal was *wrong* to murder his victim. As these examples illustrate, normative claims are ubiquitous. They inform our most unexceptional as well as our most dire activities.

Despite their pervasiveness, however, normative claims are rather mysterious. They purport to have a certain authority over us: they tell us how we ought to live, or which actions we should perform, or which ends to pursue. But what justifies this authority? What makes it the case that we should keep promises, hold doors, or go to the gym? More momentously—what makes it the case that murder is wrong?

In its general form, this is the foundational question in ethics: how is the authority of normative claims to be justified? Recently, a great deal of attention has been directed at the idea that we might answer this foundational question by turning to the philosophy of action. According to a view that I will call *constitutivism*, action has a certain structural feature—a constitutive aim—that both constitutes events as actions and generates a standard of assessment for action. We can use this standard of assessment to derive normative claims. In short, the authority of certain normative claims arises from the bare fact that we are agents.

Thus, the great hope of constitutivism is that an investigation of the structure of agency will enable us to answer the foundational question in ethics. It will reveal why certain normative claims are justified. To see how this might work, consider an example. If you understand the nature of a game, such as chess, you thereby understand a host of normative claims that regulate chess players. For example, part of what it is to play chess is to aim at checkmating your opponent. This aim simply must be present in order for a series of movements to count as an episode of chess-playing. This aim therefore seems to generate a standard of assessment for chess players: a player is successful to the extent that she fulfills the aim. Moreover, this aim generates reasons for action: if the player sees that she can achieve checkmate by moving her rook to a certain space, then she has a reason to do so.

Constitutivists call aims of this sort *constitutive* aims. A constitutive aim is present in every instance of the activity-type that it regulates. It is present precisely because its presence is part of what constitutes the activity as an instance of its kind. Or, in plainer language: if there is an activity, and participation in this activity requires having a certain aim, then all participants in that activity are going to have the aim. They are going to have the aim because otherwise, they wouldn't be participants in the activity at all. For example, chess players have the aim of checkmate because, absent this aim, they wouldn't be chess players.

Constitutivists hope to show that action itself has a constitutive aim. If we could manage this—so the hope goes—then we would be able to derive normative conclusions from that aim. Just as we can move from the fact that chess players aim at checkmate to the claim that (for example) they have reason to capture their opponents' pieces, so too constitutivists hope that we can move from the fact that action has a constitutive aim to normative claims about what agents have reason to do. Indeed, the most ambitious versions of constitutivism attempt to show that *all* normative claims are ultimately derived from the constitutive aim of action.

In the following chapters, I will show that the attractions of constitutivism are considerable. If constitutivism works, then it will provide a way of justifying normative claims without positing irreducible normative truths or grounding norms merely in subjective, variable elements of human psychology. It will thereby avoid some central and longstanding problems in ethics.

Unfortunately, constitutivism is not well understood. Explicit defenses of the theory are relatively new, having arisen only in the past two decades.[1] Accordingly, it has been difficult to determine what the essential elements of the constitutivist framework are. The first two chapters address this point by examining constitutivism in isolation from any particular view about the nature of action. I explain what is essential to the constitutivist approach, and I show what would be necessary in order to derive normative claims from facts about the nature of action. In addition, I address a number of objections that have recently been leveled at the very possibility of a constitutivist theory. I show that these objections depend on misunderstandings of the constitutivist project, and can therefore be rebutted.

Thus, Chapters 1 and 2 explain what constitutivism is and show that a constitutivist theory is possible. Of course, it is one thing to show that constitutivism could succeed, and quite another to show that it actually does succeed. The hardest part of any constitutivist theory is developing a conception of action that is minimal enough to be independently plausible, but substantial enough to yield a constitutive aim. Chapters 3 and 4 examine David Velleman's and Christine Korsgaard's attempts to do so. While Velleman's and Korsgaard's theories are extremely insightful, I argue that they

[1] While explicit defenses of constitutivism are new, some philosophers contend that aspects of the theory—described in different terminology—are present in the work of Plato, Aristotle, and Kant (see, for example, Korsgaard 2009). I will argue that aspects of constitutivism are also present in Nietzsche's work.

succumb to a common problem: each theory can generate substantive normative results only by alternating between an excessively strong conception of agency and a much weaker conception of agency. The weaker conception of agency is all that Velleman's and Korsgaard's arguments establish, while the stronger conception is necessary for their normative conclusions to follow. Absent an argument for this stronger conception of agency, then, these theories are unsuccessful.

If constitutivism is to succeed, then we need a new conception of agency. Thus far, constitutivism has been strongly associated with Kantian theories of agency. As I explain in the following chapters, Korsgaard's version of constitutivism is avowedly Kantian, and Velleman describes his theory as "Kinda Kantian" (Velleman 2009). But despite a resurgence of interest in Kant, I think it is fair to say that many philosophers believe that Kant's moral theory fails. The Kantian theory is subject to a staggering number of objections that strike many of us as decisive: the conception of agency upon which Kant relies seems empirically implausible; the moral theory depends on the assessment of maxims, and yet the idea of a maxim is terribly obscure and imprecise; the arguments attempting to show that we are committed to the Categorical Imperative are dubious; and even if we could solve those problems, the Categorical Imperative itself seems to generate no substantive results.[2] If constitutivism were bound up with the Kantian enterprise, it would inherit all of these difficulties.

But I will show that it is not. Kant was right about this: certain rules of practical reason are constitutive of agency. We can hold on to that idea while developing it in a non-Kantian manner and grounding it in a more plausible theory of action. In pursuit of this goal, Chapters 5 and 6 articulate a conception of agency that is indebted both to contemporary empirical work on human psychology and to Nietzsche's philosophical arguments. In Chapter 5, I show that in order to account for certain empirical facts about the nature of human agency, we must reject elements of the dominant philosophical conception of reflective agency. The dominant account, which I trace to Locke and Kant, distinguishes activity and passivity in agency and treats reflective or deliberative acts as paradigmatic cases of agential activity. I argue that although we need a distinction between the active and the passive in action, philosophical and empirical considerations show that this distinction has nothing to do with whether the action was brought about in a reflective or deliberative manner. I defend a new account of agential activity, according to which an agent is active in the production of her action iff two conditions are met: (i) the agent approves of her action, and (ii) further knowledge of the motives figuring in the etiology of this action would not undermine her approval of the action. By drawing on a psychologically realistic account of motivation and agency, we can show that agents constitutively aim at this form of agential activity.

[2] I will explore the first objection in Chapter 5. There is a vast literature on the other three objections. For helpful introductions to these disputes, see for example Brewer (2002) on the idea of maxims, Williams (1986) on our alleged commitment to the Categorical Imperative, and Wood (1990) on Hegel's argument that the Categorical Imperative generates no content.

Thus, Chapter 5 argues that action has a constitutive aim. But this aim, on its own, generates very little normative content. Its importance becomes apparent only when we link it to another aspect of agency, which I discuss in Chapter 6.

To bring out this second aspect of agency, Chapter 6 turns to a largely untapped source of ideas about the relationship between agency and value: the work of Nietzsche. Nietzsche might seem an unpromising source for ideas conducive to the defense of constitutivism; after all, he is famously skeptical of ethical theorizing, and he flatly denies that there are any objective facts about what is valuable. However, as I will argue in Chapter 6, Nietzsche does offer ethical ideals of his own, and his critiques of traditional morality rely on the idea that a certain value—power, in particular—has a privileged normative status. I will suggest a novel way of interpreting Nietzsche's claims: power has a privileged normative status precisely because we are committed to this value merely in virtue of acting. Nietzsche's obscure claim that all actions manifest, and are to be evaluated in terms of, "will to power" can be read as an attempt to move from a claim about the essential nature of action to a claim about value. Thus, surprising as it may seem, I will argue that we can use a Nietzschean claim about the constitutive features of action to derive a standard of success for action.[3]

In short, we can use Nietzschean considerations to show that action has a second constitutive aim: power (this term is given a special technical sense, as I will explain below). In defending this idea, I begin by considering Nietzsche's baffling claims about revaluation. Nietzsche famously argues that we must "revalue" our values, critiquing them and in some cases replacing them with new values. I argue that Nietzsche's revaluations are based upon the idea that power has a privileged normative status: power is the one value in terms of which all others values are to be assessed. If this is the correct interpretation of Nietzsche's ethical theory, though, it raises a question: how could power have a privileged status, given that Nietzsche denies that there are any objective facts about what is valuable? I argue that Nietzsche's account of agency provides the answer: he grounds power's privileged status in facts about the nature of human motivation. In particular, Nietzsche's account of *drives* entails that human beings are ineluctably committed to valuing power. So Nietzsche's ethical theory follows from his account of the nature of agency.

[3] Even the most casual readers of Nietzsche recognize that his texts are enormously complex and ambiguous. Decisively establishing that Nietzsche held any particular ethical view is no easy task; it requires sustained textual analysis, reconciliation of apparently conflicting passages, reconstruction of often fragmentary arguments, and so on. My primary goal in this volume is to defend a version of constitutivism, rather than an interpretation of Nietzsche's texts. Accordingly, I will bracket many interpretive issues in the following chapters. I will argue that some pervasive and central ideas in Nietzsche's texts seem deeply and obviously inconsistent, but make perfectly good sense if we interpret Nietzsche as a constitutivist. This strongly suggests that Nietzsche intends his ethical theory to be interpreted along constitutivist lines (though, obviously, he does not use this terminology). However, providing sufficient textual analysis to establish that this is Nietzsche's actual view would take us too far afield. Accordingly, skeptical readers can treat my reading as a way of developing some of Nietzsche's central ideas, rather than as an explication of Nietzsche's actual view. I return to these matters in Chapter 6 and in the Appendix.

Thus, if we accept a Nietzschean account of agency—as I argue that we should—then power turns out to be a constitutive aim of action. But what exactly is "power"? We might suppose that valuing power denotes valuing conquest, mastery of others, and so forth. But this is not what Nietzsche means. Power is a term of art, for Nietzsche; he gives it a special sense. To will power is to aim at encountering and overcoming resistance in the course of pursuing other, more determinate ends. In other words, to say that we will power is to say that whenever we will an end, we aim not merely to achieve the end, but also to encounter and overcome resistances that arise in the pursuit of the end. For example, to say that I will power in the pursuit of writing this book is to say that I will not only to complete the book, but also to encounter and overcome challenges or resistances in the course of doing so.

The Nietzschean account of agency entails that *all* actions have this structure: whenever we will any determinate end at all, we also will to encounter and overcome resistance in the course of pursuing that end. Although this claim is paradoxical, I hope to show that it is supported both by compelling philosophical arguments and by recent empirical work on human psychology.

Thus, I will use a Nietzschean theory of action to argue for a bipartite constitutivist theory: our actions aim both at agential activity and at power. In Chapters 7 and 8, I argue that this theory generates a range of substantive normative claims. I show that while some of these claims conform to our ordinary thoughts about what there is reason to do, others are quite surprising. In particular, this constitutivist theory requires a reassessment of some of our most cherished values, such as the positive valuation that we place on certain forms of egalitarianism and the negative value that we place on certain forms of pain. I close, in Chapter 9, by discussing the advantages that this Nietzschean version of constitutivism enjoys over competing ethical theories.

The hope, then, is that by drawing on a roughly Nietzschean theory of agency, we can answer the foundational question in ethics, showing how normative claims are justified. In particular, we can justify normative claims by showing that every agent aims jointly at activity and power.

1

Three Challenges for Ethical Theory

The most gripping and persistent philosophical problems arise when we have strong and unshakeable convictions that certain claims must be true, and yet, upon reflection, we cannot see how they could so much as aspire to truth. So, freedom is a philosophical problem because our practices and the experiences of deliberation firmly wed us to the idea that we must have a distinct form of control over our own actions, and yet a realistic view of the world seems to commit us to seeing all of our actions as determined by events not under our control. Consciousness is a philosophical problem because our understanding of the ways in which our brains process stimuli is ever increasing, and yet we want to say that these results still leave it utterly mysterious why some of these processes are accompanied by conscious awareness.[1] And, turning now to our topic, morality is a philosophical problem because we want to say that there are universally valid normative facts, and yet we cannot see how such facts could be woven into the fabric of the universe. We want to say that needlessly inflicting suffering on innocents is wrong, universally wrong, for all people and all times; and yet even the most appalling torments do not have inscribed on them "you shall avoid me."

At the close of the eighteenth century, it was still possible for Kant to claim in all earnestness that "two things fill the mind with ever new increasing wonder and awe, the oftener and more steadily we reflect on them: the starry heavens above me and the moral law within me" (*Critique of Practical Reason*, 5:161). Today, many find that steady reflection on morality inspires less wonder and awe than skepticism and detachment. A host of studies purport to show that moral beliefs are relics of affects and dispositions instilled deep in our evolutionary past; others argue that morality is a vestige of religion; and philosophers such as Nietzsche predict that we will unlearn our awe of morality, just as past ages unlearnt their awe of astrology and alchemy (HH 4, BGE 32 and 188).[2]

[1] Huxley and Youmans wrote: "what consciousness is, we know not; and how it is that any thing so remarkable as a state of consciousness comes about as the result of irritating nervous tissue, is just as unaccountable as the appearance of the Djin when Aladdin rubbed his lamp in the story" (Huxley and Youmans 1868, 178).

[2] For an excellent overview of the evolutionary arguments, see Joyce (2006, Chapter 6). For an example of the religious arguments, see Anscombe (1958), Weber (2002), or virtually any of Nietzsche's works.

What sparks this skepticism about morality? Why have many reflective individuals gone from seeing in morality a source of awe to finding it dubious and archaic? Doubtless the reasons are manifold, but in the following pages I will trace one answer. An adequate account of morality would have to overcome three challenges: it would have to show why we should have confidence in our moral beliefs, why these moral beliefs don't rely on outmoded or outlandish metaphysical claims, and why we should take morality to be prescriptive. I call these the epistemological, metaphysical, and practical challenges. Versions of these challenges are familiar in the literature on ethics, but I will argue that Nietzsche presents especially powerful forms of each challenge. Moreover, I will argue that the chief ethical theories encounter serious difficulties in trying to overcome these challenges. It is the perceived inability to meet these challenges that leads many individuals to moral skepticism. But I will also argue that this result is not inevitable: we can avoid skepticism by pursuing a different strategy in ethics—constitutivism.

1. Three challenges for ethical theory

Before beginning, a word on what morality is. Moral claims are normative. They purport to have a certain authority over us. They purport to be claims according to which we should regulate or guide our actions. But not all normative claims are classified as moral claims. The claim "you should drink coffee" is not typically taken as moral; the claim "you should not murder innocent children" is. So what distinguishes moral claims from non-moral normative claims?

The answer to this question is not obvious. However, many philosophers believe that there is at least one necessary condition for a normative claim's being a moral claim: *universality*. Moral claims are *universal* in the sense that they apply to all agents. We typically take only some agents to have reason to drink coffee, but we hold that all agents should refrain from murdering innocent children. The claim about drinking coffee has a suppressed premise: *if* you enjoy the taste of coffee (or want to wake up, or . . .), then you should drink coffee. The claim about murder does not; it purports to apply to *all* agents, regardless of facts about their preferences, goals, and characters. After all, we do not take the fact that someone has a desire to murder as undermining the authority of the claim that murder is wrong.[3]

[3] For this reason, some philosophers take moral claims to be *categorical*. A normative claim is categorical if it applies to agents regardless of their preferences, goals, or aims. Notice that a normative claim can be universal without being categorical: for example, if a normative claim applies to agents in light of their having a particular aim, and if this particular aim is present in all agents, the normative claim will be universal but not categorical. I won't be assuming that moral claims are categorical, though I discuss the possibility both below and in Chapter 4. In addition, it is worth noting that some philosophers argue that moral claims have yet a third distinguishing feature: *overridingness*. That is, moral claims either always or typically trump competing normative claims: if I am faced with a conflict between fulfilling a moral demand and some other demand, I am obligated to fulfill the moral demand. This feature is controversial; many philosophers offer powerful arguments against it (Sidgwick 1981 is a classic example). I will not assume that moral claims are overriding.

Universality alone may not be enough to distinguish moral claims from other normative claims. Consider the claim "you ought to regulate your beliefs in accordance with the relevant evidence." That's a plausible candidate for a universal normative claim, but it would be unusual to classify it as a moral claim. So we might need to say something about the *content* of universal normative claims. Here, there are several options. We might think that moral claims govern our interactions with other agents; paradigmatic moral claims might then be prohibitions on harming others, requirements to aid others, and so forth. Or we might think that moral claims specify what it is to live well or to flourish; paradigmatic moral claims might then tell us to seek happiness, or achievement, or to avoid squandering our capacities.

I think that any recognizably moral claim will have some such content; it will either govern our relationships with other agents or specify what it is to live well. However, I do not want to assume, in advance, that moral claims have any *specific* content. I will not assume, for example, that we can justify a universal normative claim requiring us to help others, or to be compassionate, or to cultivate our talents. So I'll start with a rather loose definition of morality: moral claims are universal normative claims that either specify appropriate behavior toward other agents or specify what it is to live well. We'll refine this notion as we progress.

Below, I will consider three challenges for morality. That is, I will consider three challenges to the very idea that there can be universal normative claims specifying appropriate behavior toward others or specifying what it is to live well. I will then ask whether the dominant ethical theories give us a way of answering these challenges.

1.1 The epistemological challenge

The first challenge for morality arises from a simple fact: morality has a history.[4] To illustrate the relevance of this point, let's consider Nietzsche's *On the Genealogy of Morality*. As is well known, the *Genealogy* sets out to demonstrate that many of our most basic moral beliefs arose approximately two thousand years ago as the product of a resentment-inspired revolt carried out by a lackluster, vengeful underclass. Put briefly, Nietzsche's argument is as follows. In the ancient world, the dominant moral code was organized around a good/bad dichotomy, where the traits labeled "good" were those associated with the nobility, and those labeled "bad" were those associated with the commoners. Strength, self-assertion, power, desire to rule, competition, wealth,

[4] It is important to notice that the term "morality" can be used in either a descriptive or a normative sense. Descriptively, the term refers to the moral code that is generally accepted within a particular society or social group at a particular time. Thus, we might speak of the morality of the fifth-century Athenians, the morality of the antebellum South, and so on. Normatively, the term "morality" refers to a putatively *correct* moral code. These can come apart: while the fifth-century Athenians accepted moral claims such as "slavery is permissible," one hopes that this was an error; one hopes that they ought to have accepted the claim that slavery is impermissible. When I say, above, that morality has a history, I am starting with the *descriptive* sense of morality.

health, and beauty were taken as good; weakness, humility, lack of power, servility, inability to compete, ordinariness, and ugliness were taken as bad. Nietzsche argues that this system of evaluations was, for a time, accepted by the bulk of society: slaves, commoners, and nobles all embraced these values. About two thousand years ago, Nietzsche claims, things began to shift. A new set of values emerged. These new values—which Nietzsche calls the "good/evil" or "slave" morality—invert many of the earlier values. Thus, manifestations of strength, self-assertion, power, desire to rule, competition, wealth, and certain forms of beauty are taken as bad; weakness, meekness, humility, lack of power, servility, poverty, compassion, concern with suffering, and an idea of equality are taken as good.[5] These values are most clearly present in early versions of Christianity, but many of them remain today.[6]

To see how Nietzsche's story constitutes a challenge to moral philosophy, we must focus on three crucial claims. First, ancient and modern moralities endorse distinct and conflicting sets of values. Second, an examination of morality's history reveals that these changes in the moral code cannot be construed as mere refinements of earlier values, but must instead be seen as discontinuous breaks. Third, these breaks are best explained by psychological and social factors rather than by appreciation of rational considerations. I will explain each of these claims below.

I take it that the first claim is beyond dispute, but let me illustrate it with a non-Nietzschean example. To this end, consider Aristotle. Aristotle claims to be systematizing certain culturally pervasive intuitions (*endoxa*), rather than developing revisionary moral claims, so his writings are a particularly helpful guide to views prevalent in the ancient world. Aristotle heaps praise on a character trait that he calls megalopsychia. Perhaps the best translation for this Greek word is "greatness of soul." Megalopsychia, Aristotle tells us, is "a sort of crown of the virtues; for it makes them greater, and is not found without them" (*Nicomachean Ethics* 1124a1). What distinguishes this character trait? Aristotle remarks that the individual with megalopsychia distinguishes between persons of high and low rank (1124b15–25); he disdains the honor and praise emanating from persons of low rank, and accepts the praise of his high-ranked peers (1124a5–11); and he is motivated by honor "again, it is characteristic of the [man of megalopsychia] . . . to be sluggish and to hold back except where great honor as a great result is at stake, and to be a man of few deeds, but of great and notable ones" (1124b22–26). So, the individual with megalopsychia is acutely sensitive to social hierarchy and motivated by the desire to be perceived as honorable by those of

[5] I am eliding a complication that won't be relevant for our purposes: Nietzsche claims that slave morality replaces the concept of bad with the concept of evil. For helpful analyses of this point, see Reginster (1997) and Leiter (2002).

[6] Nietzsche's arguments for these claims are given in GM I. For helpful discussions, see for example Reginster (1997), Ridley (1998), Leiter (2002), Janaway (2007), Owen (2007), and Wallace (2007). I address these claims in more detail in Katsafanas (2011d).

high social rank. We might put the point less kindly: honor-seeking elitism is the "crown of the virtues."[7]

Clearly, modern individuals do not take these character traits to be virtues, much less the highest virtues. Indeed, our modern moral code would be more likely to deem this type of pride as a vice. The Bible is illustrative: we are told that "God opposes the proud but gives grace to the humble" (1 Peter 5:5).

Speaking of humility—Aristotle labels it a vice:

> for the unduly humble man, being worthy of good things, robs himself of what he deserves, and seems to have something bad about him from the fact that he does not think himself worthy of good things, and seems also not to know himself; else he would have desired the things he was worthy of, since these were good . . . Such a reputation, however, seems actually to make them worse; for each class of people aims at what corresponds to its worth, and these people stand back even from noble actions and undertakings, deeming themselves unworthy, and from external goods no less. (1125a19–27)

The shifting valuations attached to humility and greatness of soul are evidence in favor of Nietzsche's first claim: ancient and modern moralities endorse conflicting values.[8] Examples could be multiplied.[9]

So the first claim is uncontroversial. But it might also seem insignificant. After all, the claim that beliefs about morality have changed over time is not surprising. The same is true of every field of human inquiry. Our current beliefs about physics, chemistry, biology, and so on also grew out of earlier forms.

But this brings us to the second point: Nietzsche argues that history reveals not a smooth process of rational development, but a series of *discontinuities* between the moral beliefs embraced at different times. The later moral systems cannot be understood as rational developments of the former, but must be seen as distinct. To see what Nietzsche has in mind, notice that certain valuations can be understood as natural extensions or developments of earlier valuations. When a government moves from claiming that all property-holding white males deserve equal treatment to claiming that all persons, regardless of wealth, race, or gender, deserve equal treatment, this can be seen as a development of the internal logic of valuing equality: an ideal that is implicit or imperfectly realized in the earlier moral code is developed, rendered more consistent, and made fully explicit. Cases of this sort give us no reason for skepticism about our

[7] There is a tradition of interpreting megalopsychia in ways that make it seem less objectionable to modern sensibilities: see, for example, Crisp (2006) and Sarch (2008). With Cordner (1994), I find it preferable simply to accept that Aristotle's virtues differ from those that we would endorse. For additional reflections on the striving for something like megalopsychia in the ancient world, see Nietzsche, *Homer's Contest*.

[8] Of course, I am not claiming that the ancient and modern worlds were univocal in their respective valuations. For example, by the early modern period we can find a number of philosophers arguing that humility is not a genuine virtue, and today our attitude toward humility is decidedly mixed.

[9] Consider a few obvious examples: the valuations attached to slavery, torture, public execution, cannibalism, imperialism, monogamy, sexual promiscuity, masturbation, and homosexuality have undergone dramatic shifts over time.

values. But other values are not like this. We moved from considering elitism as good to considering it evil; we moved from considering humility bad to considering it a central virtue. It is difficult to see this as anything less than a complete inversion of these values. Nietzsche's claim is that many of the changes from classical to modern moral codes represent just such profound breaks, rather than a continuous process of rational development.[10]

Of course, it is compatible with this story that our current moral beliefs are better aligned with the truth. After all, there are also discontinuities in science: it has been clear since Kuhn (1970) that the move from Newtonian to relativistic physics cannot be understood as a process of smooth continuous development, but instead reveals discontinuities and breaks. Nevertheless, we take relativistic physics to provide a more accurate representation of the physical world than does Newtonian physics. The same could be true of our modern moral code.[11] For this reason, the presence of breaks and discontinuities does not by itself imply that the system in question is problematic.

Nietzsche is aware of this. Indeed, he claims that "there is no more important principle" for the study of history than this: "the cause of the origin of a thing and its eventual utility, its actual use and arrangement in a system of purposes, lie worlds apart; whatever exists, having somehow come into being, is again and again reinterpreted to new ends, taken over, transformed, and redirected" (GM II.12). In other words, the history of *all* systems of beliefs will display discontinuity, alogical leaps, and so forth. "The entire history of a 'thing,' an organ, a custom can in this way be a continuous chain of signs of ever new interpretations and adjustments, whose causes do not even have to be related to one another but, on the contrary, in some cases succeed and alternate with one another in a purely chance fashion" (GM II.12). Given the ubiquity of alogical developments, Nietzsche warns us against committing the genetic fallacy:

> The inquiry into the origin of our evaluations and tables of the good is in absolutely no way the same as a critique of them, as is so often believed: even though the insight into some *pudendo origo* certainly brings with it a feeling of diminution in the value of the thing that originated in that way and prepares the way to a critical mood and attitude toward it. (KSA 12:2[189]/WLN 95)[12]

As Nietzsche notes here, the discontinuities alone don't undermine our moral code.

[10] Note that this is consistent with other values being held constant. Nietzsche's claim is not that *every* valuation present in the ancient world was inverted or altered in the modern world; his claim is simply that many of the central valuations were inverted or altered.

[11] Indeed, we can put Nietzsche's point in terms familiar from the philosophy of science: changes in moral codes resemble paradigm shifts rather than normal science.

[12] Compare GS 345: "Even if a morality had grown out of error, this would not so much as touch on the problem of its value. . . . The mistake made by the more refined among them [historians of morality] is that they uncover and criticize the perhaps foolish opinions of a people about their morality, or of humanity about all human morality—opinions about its origin, religious sanction, the superstition of free will, and things of that sort—and then suppose that they have criticized the morality itself."

The discontinuities revealed by history do, however, prompt us to ask *why* these shifts occurred. As Nietzsche puts it, history makes us *feel* that the value of the thing in question has diminished, and thereby "prepares the way to a critical mood and attitude toward it." In short, these histories make us wonder why we hold the moral beliefs that we do.

With shifts in the sciences, we at least have some conception of what motivated changes in theories and what would serve as a check on an incorrect theory: we can appeal to explanatory adequacy, coherence, simplicity, and so forth. There are well-known problems with these standards, and they clearly won't be sufficient conditions for deciding between competing theories, but we at least have a sense of what we are after.[13] With morality, though, what are we after? Conformity to moral intuitions?[14] Conformity to emotions such as compassion or sympathy? The promotion of some end, such as long-term self-interest or happiness? It is not obvious: each of these purported standards of success is highly controversial—*far* more controversial than in the scientific case. Indeed, each of these standards would be rejected by proponents of certain moral codes.[15] Below, we'll see that Nietzsche calls all of these grounds into question.[16,17]

If we had a theory-neutral criterion of success, then discontinuities in moral codes would not be troubling. For we could use the criterion to determine whether, despite the alogical leaps, the successive moral codes were getting closer to or further from the truth. But we don't have that—unless one of the moral theories that I will discuss below can provide it.

[13] For some classic discussions of the problems with these standards, see, for example, Popper (1959), Kuhn (1970), and Laudan (1977).

[14] Some philosophers do try to draw analogies between the role of data in science and in morality. But, as Peter Singer notes, this analogy is at best highly strained: "The analogy between the role of a normative moral theory and a scientific theory is fundamentally misconceived. A scientific theory seeks to explain the existence of data that are about a world 'out there' that we are trying to explain. Granted, the data may have been affected by errors in measurement or interpretation, but unless we can give some account of what the errors might have been, it is not up to us to choose or reject the observations. A normative ethical theory, however, is not trying to explain our common moral intuitions. It might reject all of them, and still be superior to other normative theories that better matched our moral judgments . . . " (Singer 2005, 345).

[15] We might think that we could avoid this problem by claiming that morality aims at human flourishing. But as soon as the notion of flourishing is given any substantive content, this claim becomes controversial. I address this point below and, in a rather different way, in Katsafanas (2011d).

[16] Thus, in *Daybreak* 106 Nietzsche remarks that morality is presented as enhancing or preserving mankind. But "preservation of *what?* Is the question one immediately has to ask. Advancement *to what?* is the essential thing—the answer to this *of what?* and *to what?* not precisely what is left out of the formula?" He goes on to note that there are many potentially conflicting goals here.

[17] Philosophers sometimes appeal to convergence in moral beliefs in order to support the idea that we're moving increasingly closer to a correct moral theory. Mere convergence won't be convincing, though, unless we have some reason for thinking that we aren't converging toward errors. After all, from late antiquity to the Middle Ages, we can see the moral codes of various European nations as converging toward Judeo-Christian values, and yet we can give an obvious explanation for this convergence: the spread and increasing political and cultural influence of Christianity.

Nietzsche's third point becomes relevant here. If these changes in moral codes didn't occur in response to rational reflection, what prompted them? Nietzsche argues that attention to history reveals that moral systems arose and changed for *psychological and social* reasons. The ancient nobility, Nietzsche tells us, affirmed their own way of life and hence deemed it good.[18] Their positions of authority enabled them to promulgate these values. The fundamentals of our modern moral code arose when a vengeful underclass rejected the earlier moral system; consequently, it rests on a psychological state that Nietzsche calls *ressentiment*.[19] It took hold for social reasons: a set of priestly figures preached values that would appeal to the oppressed, downtrodden servant classes, who constituted the bulk of society.[20] Thus, rather than arising in response to appreciation of rational considerations, moral shifts occurred for contingent psychological and social reasons. If this is correct, it should undermine our confidence in them.

Aspects of Nietzsche's story may be fanciful, and at the very least we would need more evidence for these claims than Nietzsche himself provides. But we can set that aside; the details are irrelevant for our purposes. What matters is the truth of a general claim: there are discontinuities in the development of moral codes, and these discontinuities are best explained by psychological and social considerations. Do we have reason to believe this is true?

I think we do. Even if Nietzsche's evidence for this claim is rather spare, it does not stand alone. We have Marx's arguments that moral shifts are best explained by economic factors.[21] We have Weber's arguments linking some of our central moral beliefs to religious assumptions.[22] We have evidence from anthropology and evolutionary biology that certain moral beliefs arose in response to highly contingent and now vanished circumstances, such as conditions of low population density or the authority of various religions.[23] And there are simpler ways of making the same point. Gilbert Harman gives a very nice example. Most cultures seem to accept a judgment of the following form: *harming someone is worse than failing to help someone*. For example, if I kill someone, this is terrible; but if I know that an individual will starve to death unless I donate a negligible amount of money that I would otherwise waste on frivolous entertainments, and nonetheless fail to donate the money, this is widely regarded as perfectly acceptable. This is rather odd: the consequences are the same, after all. There is a voluminous literature investigating potential justifications for the claim. But Harman points out that we can give an exceedingly simple and elegant

[18] "The noble type of man experiences *itself* as determining values . . . Everything it knows as part of itself it honors: such a morality is self-glorification" (BGE 260).

[19] See GM I.7–11. "The slave revolt in morality begins when *ressentiment* becomes creative and gives birth to values" (GM I.10). *Ressentiment* is Nietzsche's term for a vengeful hatred born of impotence; see Reginster (1997), May (1999), and Wallace (2007) for discussions.

[20] For discussions of this point, see the *Genealogy* and the *Antichrist*.

[21] For introductions to these ideas, see Cohen (2001) and Wolff (2002).

[22] See, for example, Weber (2002).

[23] See Prinz (2007, 220–87) for a helpful overview. See also Joyce (2006) and Sinnott-Armstrong (2007).

explanation for why this valuation became widespread: it aids the wealthy and powerful.

> Whereas everyone would benefit equally from a conventional practice of trying not to harm each other, some people would benefit considerably more than others from a convention to help those who needed help. The rich and powerful do not need much help and are often in the best position to give it; so, if a strong principle of mutual aid were adopted, they would gain little and lose a great deal, because they would end up doing most of the helping and would receive little in return. On the other hand, the poor and the weak might refuse to agree to a principle of non-interference or noninjury unless they also reached some agreement on mutual aid. We would therefore expect a compromise . . . the expected compromise would involve a strong principle of noninjury and a much weaker principle of mutual aid—which is just what we now have. (Harman 1977, 111)

In this passage, Harman provides a mini genealogy of one moral principle. Just as with Nietzsche, Harman's story doesn't show that the moral principle is false. But it does make us wonder whether any good reasons can be given for its acceptance. In general, if we have a powerful social or psychological explanation for why we hold a value, and we have difficulty seeing what independent grounds can be given for the value's acceptance, then our confidence in the value should be undermined.

To sharpen this Nietzschean argument, it will be helpful to contrast it with the traditional argument from disagreement. John Mackie gives a classic statement of that argument:

M1. There is moral disagreement: different cultures exhibit different moral beliefs.
M2. The best explanation for this disagreement is that there are no objective facts about morality.[24]
M3. Therefore, there are no objective facts about morality.[25]

This argument from disagreement has come under criticism.

Some critics have objected that there really isn't so much disagreement about morality. In particular, apparent disagreements about values often turn out to be based on factual disagreements or ignorance. For example, consider the claim that slavery is permissible. Some individuals who endorsed this judgment supported it with erroneous factual claims. Aristotle claimed that certain individuals had physical and psychological aspects that rendered them "natural slaves";[26] analogously, some

[24] Mackie writes, "the argument from relativity has some force simply because the actual variations in the moral codes are more readily explained by the hypothesis that they reflect ways of life than by the hypothesis that they express perceptions, most of them seriously inadequate and badly distorted, of objective values" (Mackie 1977, 37).

[25] Loeb (1998) offers a different version: if moral realism is correct, then moral questions must be in principle resolvable; but a number of moral questions are not even in principle resolvable; so moral realism is false. See Leiter (forthcoming b) for a version of this argument that focuses on putatively irresolvable disagreement about moral *theories* rather than moral *beliefs*.

[26] Aristotle asks "is there any one thus intended by nature to be a slave, and for whom such a condition is expedient and right"? He answers, "There is no difficulty in answering this question, on grounds both of

American slaveholders claimed that slaves were cognitively inferior and therefore were best served by slavery. It is possible that if these false factual beliefs had been corrected, then Aristotle and some of the American slaveholders would have abandoned their belief that slavery was justified. Generalizing this point, some philosophers respond to Mackie's argument by rejecting M2: these philosophers claim that if we get all of our non-moral facts straight, moral disagreement will vanish.

Another explanation for moral disagreement is simply that morality is hard. After all, there has also been a great deal of disagreement about physics, biology, economics, and so forth. We don't readily conclude, from the fact that different cultures or different times have disagreed about the nature of physical reality, that there are no objective facts about physics. We don't readily conclude, from the fact that economists disagree about which tax policy would maximize GDP, that there is no fact of the matter about which tax policy would maximize GDP. Morality might have an analogous explanation: it is complex and difficult.

Moreover, in morality as in economics there are clear pressures toward self-deception. When we consider a question such as "do low tax rates on the wealthy maximize GDP?," the wealthy have a clear stake that might bias their answers. Just so, when we consider certain moral questions, agents have clear stakes that might bias their answers. Consider one example: the ancient Greeks and the early American slaveholders were under pressure to regard slavery as permissible because it was economically advantageous for those involved in certain modes of agriculture. These pressures could explain why disagreement arises.

In short, critics allege that the best explanation for moral disagreement is a combination of factual disagreement, difficulty of the subject matter, and pressures toward self-deception. These criticisms have made the traditional argument from disagreement seem less persuasive. But notice that Nietzsche's argument is very different. We can start with Mackie's M1:

M1. There is moral disagreement: different cultures exhibit different moral beliefs.

From there, Nietzsche's argument takes a different turn:

N2. These differences in moral beliefs were caused by social and psychological factors.
N3. In general, the kinds of social and psychological factors that shift moral beliefs do not track the truth.
N4. If we recognize that a belief was caused by factors that do not track the truth, then we need justification for continuing to hold this belief.
N5. Therefore, we need justification for continuing to hold our moral beliefs.[27]

reason and of fact. For that some should rule others and not be ruled is a thing not only necessary, but expedient; from the hour of birth, some are marked out for subjection, others for rule" (*Politics* 1254a).

[27] For a related reading of Nietzsche's argument, see Sinhababu (2007, 276–9). Sinhababu provides a helpful discussion of the ways in which the presence of unreliable psychological processes in belief-formation

If Nietzsche is right about N2, then claims about the difficulty of morality will be irrelevant. And indeed, the responses to Mackie's argument are committed to some version of N2: the respondents want to show that certain moral beliefs, which they regard as disagreeable (such as Aristotle on slavery) were explained by social and psychological factors, whereas others (such as all of ours) were not. Once we've admitted the possibility of influence, though, we need some reason for thinking that it does not affect our own moral beliefs.

So M1 and N2 seem well supported. We can hardly deny N3. No one is going to argue that resentment, desire for power, desire for economic advantage, and so forth are psychological mechanisms that track the truth. N4 also seems uncontroversial. For these reasons, Nietzsche's argument won't be defused by the traditional objections to Mackie's argument from disagreement.[28] Nietzsche's argument poses a genuine challenge for moral philosophy.[29]

With these points in mind, we can summarize the epistemological challenge as follows: attention to the way in which morality developed undermines our confidence in and justification for our current evaluations. If our commitment to all of our values can be explained in the above fashion, then the worry is that we will not be able to sustain our commitments. After all, morality is *demanding*. It tells us how to live. It tells us how to structure our interpersonal relationships. Kant claims that it aspires to overrule any competing inclinations: the thought of moral duty "strikes down all *arrogance* as well as vain *self-love*" (Kant, *Critique of Practical Reason*, 5:86). All of this is right. The question is whether morality can continue to occupy these roles once we appreciate its origins. Nietzsche thinks not: he tells us that

> If you had thought more subtly, observed better, and learned more, you certainly would not go on calling this 'duty' of yours and this 'conscience' of yours duty and conscience. Your knowledge of the way in which moral judgments have originated would spoil these grand words for you, just as other grand words, like 'sin' and 'salvation of the soul' and 'redemption' have been spoiled for you. (GS 335)

Below, we will ask whether Nietzsche is right.

should undermine our faith in the belief. See also Leiter, who writes, "we *should* be suspicious of the epistemic status of beliefs that have the wrong causal etiology" (2006, 104).

[28] Enoch (2009) attempts to rebut all forms of the argument from disagreement. He considers ten versions of the argument and offers responses to each. Despite Enoch's aspiration to comprehensiveness, however, he does not consider anything like the Nietzschean argument from disagreement mentioned above.

[29] Nietzsche's argumentative strategy is clearly presented in the following passage: "*Historical refutation as the definitive refutation*.—In former times, one sought to prove that there is no God—today one indicates how the belief that there is a God could *arise* and how this belief acquired its weight and importance: a counter-proof that there is no God thereby becomes superfluous.—When in former times one had refuted the 'proofs of the existence of God' put forward, there always remained the doubt whether better proofs might not be adduced than those just refuted: in those days, atheists did not know how to make a clean sweep" (D 95). In this passage, Nietzsche indicates that seeing how the belief in God originated undermines the belief—not because it shows that the belief is false, but because it shows that the belief arose for dubious reasons.

1.2 The metaphysical challenge

So far, we have one criterion of adequacy for an ethical theory: it must explain why, despite the discontinuities in moral beliefs and psychological explanations for these discontinuities, we should have confidence in our current moral beliefs. This brings us to the second challenge. To introduce the challenge, it is helpful to begin with an objection. When speaking to educated, non-religious individuals outside of philosophy departments, there is a very common reaction to claims about morality: the idea of universal values is antediluvian, a relic of discredited religious or outmoded scientific accounts of the world. The idea that there are any objective facts about what we *should* do, or what is *valuable*, is just one last form of anthropocentrism lurking among the scientifically ignorant. A realistic, empirically informed account of morality shows it to be *nothing more* than a series of conventions and customs, devoid of any deeper justification.

Why might the idea of universal values seem outmoded? The problem is one of *naturalism*. We want a theory that is compatible with our best account of the natural world, and morality seems to face two problems on this score.

First, some attempts to justify conventional morality appeal to properties that seem fanciful. Again, John Mackie gives a classic formulation of this objection: "if there were objective values then they would be entities or relations of a very strange sort, utterly different than anything else in the universe" (1977, 38). For if such values existed, then it would be possible for a certain state of affairs to have "a demand for such-and-such an action somehow built into it" (1977, 40). And this, Mackie concludes, would be a decidedly "queer" property.

Of course, there are controversies regarding which qualities should count as queer. Contemporary physics posits a number of bizarre properties. But consider just how odd moral facts would be: they would be facts with intrinsic prescriptivity or imperativeness. A moral fact would be something that directs us or commands us to act in a certain way.[30] This is what makes purported moral facts queer: they *demand* that we do something, independently of any facts about our motives or goals. And it's hard to see what kind of facts or properties could have this intrinsic demandingness built into them. Richard Garner puts the point well: "it is hard to believe in objective prescriptivity because it is hard to make sense of a demand without a demander, and hard to find a place for demands or demanders apart from human interests and conventions. We know what it is for our friends, our job, and our projects to make demands on us, but we do not know what it is for *reality* to do so" (Garner 1990, 143). In short, what's queer is the idea of intrinsic prescriptivity lodged in the world.

[30] Philosophers sometimes express this point in terms of motivation: appreciation of a moral fact is supposed to be capable of motivating the agent. For now, I want to set aside questions about whether morality necessarily motivates, and focus instead on the fact that a *demand* for motivation is somehow built into the moral fact. I'll address the question about whether moral facts are necessarily motivating in the next section.

This brings us to the second way in which moral philosophy can seem to run up against naturalistic troubles: certain attempts to justify conventional morality make presuppositions about human beings that are either demonstrably false or otherwise problematic. For example, Aristotle's moral theory relies on an outmoded natural teleology that implies that human beings have a function. Kant is committed both to the idea that we can draw a clean distinction between reason and passion, and that all actions are performed on maxims. Contemporary biology and psychology give us reason to doubt each of these claims.[31] Insofar as the justifications of moral theories require or presuppose indefensible claims about human beings, they are unacceptable.

In sum, then, we have two requirements on an adequate moral theory. First, the theory must be metaphysically respectable: the account of reasons and values must not appeal to any non-natural qualities. Second, the theory must be psychologically realistic: the account of reasons and values must not presuppose a model of agency or human psychology that is ruled out by our best philosophical and scientific accounts.

The psychological requirement seems to me the more difficult one. We will see below that most of the dominant ethical theories can avoid appeal to non-natural qualities, and thereby meet the metaphysical constraint. However, the psychological constraint is a substantial impediment.

Nietzsche is again relevant here, for he is focused mainly on the psychological point.[32] He takes it for granted that we should avoid metaphysically extravagant properties, dismissing views that posit "intercourse between imaginary beings" and

[31] For criticisms of Aristotle along these lines, see Williams (1986, Chapter 3); for Kant, see Blackburn (2001, Chapter 8), Leiter (2002), and Risse (2007).

[32] Nietzsche endorses some version of naturalism, but it is not obvious *which* version. Brian Leiter has argued that Nietzsche is a "methodological naturalist"; that is, Nietzsche thinks "philosophical inquiry . . . should be continuous with empirical inquiry in the sciences" (Leiter 2002, 5). He adds that Nietzsche is best viewed as a "*speculative* methodological naturalist" (2002, 5, emphasis added). Speculative naturalists do not merely take the current scientific discourse to be correct, but aim to go further; they "construct theories that are 'modeled' on the sciences . . . in that they take over from science the idea that natural phenomena have deterministic causes" (Leiter 2002, 5). This seems to me entirely correct, but it does leave open the difficult question of what counts as "modeling" a philosophical theory on the sciences.

There is no denying that Nietzsche was fascinated with the sciences of his day; he writes, in *Ecce Homo*, "A truly burning thirst took hold of me: henceforth I really pursued nothing *more* than physiology, medicine and natural sciences" (EH III: HH-3). But, at the same time, he criticizes much of the science that was current in his time, dismissing the "clumsy materialists" who "can hardly touch on the soul without immediately losing it" and inveighing against "materialistic atomism" (BGE 12). One important aspect of this critique is that unlike some naturalists, Nietzsche has no aspirations to eliminate all evaluative discourse. He thinks that affective and purposive orientations toward the world already include evaluations; indeed, he writes that even "sense-perceptions are permeated with values" (KSA 12:2[95]). He seems untroubled by the idea that certain psychological descriptions must be posed in evaluative terms. In short, Nietzsche's accounts of the natural are complex. As Janaway puts it,

> If Nietzsche's causal explanations of our moral values are naturalistic, they are so in a sense which includes within the "natural" not merely the psychophysical constitution of the individual whose values are up for explanation, but also many complex cultural phenomena and the psychophysical states of past individuals and projected types of individual. (Janaway 2007, 53)

But, as Leiter has argued, this is consistent with the idea that Nietzsche is a naturalist (Leiter forthcoming a). Nietzsche wants "to complete our de-deification of nature . . . [and] to 'naturalize' humanity in terms of a

rely on "an imaginary natural science (anthropocentric; no trace of any concept of natural causes)" (A 15; italics removed). But he seems to think the real work is done by the psychological constraint: the models of agency, consciousness, deliberation, and knowledge employed by the traditional ethical theories are problematic. Thus, a very common form of objection in Nietzsche's works is this: Plato, Kant, Mill, or some other philosopher has a defective, unrealistic account of agency; recognizing this fact vitiates the philosopher's moral theory.[33]

The psychological constraint *seems* simple enough: as Nietzsche puts it, all that is required is that we "translate man back into nature" (BGE 230). The idea that we need to translate *back* implies that our current conception of human beings has somehow gone astray. For example, Nietzsche writes, "We no longer derive man from 'the spirit' or 'the deity'; we have placed him back among the animals . . . Descartes was the first to have dared, with admirable boldness, to understand the animal as machine. The whole of our physiology endeavors to prove this claim. And we are consistent enough not to except man, as Descartes still did . . . " (A 14). Nietzsche's idea, then, is that our concept of *human being* and *human agent* must be freed from the accretions of defunct religious and philosophical interpretations. But these accretions and errors aren't obvious. Almost everyone agrees that we should avoid appeal to psychologically unrealistic accounts of agency; almost everyone disagrees about what counts as psychologically unrealistic. As a result, uncovering the psychological errors inherent in certain moral theories is a substantial task—and one that will occupy much of Chapters 3 through 6.

1.3 The practical challenge

So far, we have two challenges for ethical theory. First, there is an epistemological challenge: an adequate ethical theory must explain why we should have confidence in our moral beliefs. Second, there is a metaphysical challenge: the theory must be naturalistically respectable, both in its treatment of normative properties and its analysis of agency. This brings us to a third challenge: as mentioned above, morality is prescriptive. It not only tells us what to do, but purports to outweigh many competing claims about what to do. An adequate ethical theory must explain why and how morality has this grip on us.

A common way of making this point is by appealing to Motivational Judgment Internalism (hereafter MJI): if an agent judges that she ought to ϕ, then insofar as she is rational she is motivated to ϕ.[34] The idea behind MJI is quite simple: if you are rational,

pure, newly discovered, newly redeemed nature" (GS 256). In the following chapters, I will explore how these considerations inform Nietzsche's account of motivation and agency.

[33] For a few examples, see BGE 32, GM I.13, GM II.2, and TI VI. We will examine these claims in depth in the following chapters. For discussions of these claims, see Leiter (2002), Leiter and Sinhababu (2007), Risse (2007), and Gemes and May (2009).

[34] Scanlon has a clear statement of this claim: "If a person judges that she has conclusive reason to do *X* at *t*, then two things follow. First, insofar as she does not abandon or forget this judgment, she is irrational if she does not intend to do *X* at *t*. Second, the fact that she holds this judgment about reasons can explain her

then the answers to the questions "What ought I to do?" and "What will I do?" are the same. For, as Ralph Wedgwood puts it, "if you are rational, your question 'What ought I to do?' *is* a deliberative question about what to do" (2007, 25). So, if I am rational, when I judge that I ought to brush my teeth, I will acquire some motivation to do so. We can see this as a constraint on the content of normative claims: in order for it to be true that I ought to ϕ, the thought of ϕ-ing must be capable of motivating me.

Although appeals to MJI were at one point quite common in the literature, a number of objections have emerged. Some philosophers have argued that once we include the caveat that MJI applies only to rational agents, MJI becomes stipulative or merely definitional. Others have argued that agents can be amoralists, who make moral judgments perfectly well but are not motivated to conform to them.[35] If this is correct, then the connection between moral judgment and motivation drawn by MJI may be too tight.

However, we needn't resolve these disputes, for there is a second way of putting the point. Our moral beliefs have a grip on us. Morality tells us what to do. We may not do what it tells us to do; if amoralists are a real possibility, we may not even be motivated in the slightest to do what it tells us to do. But if morality could be completely severed from motivation—if my judgments about what is valuable, what is wrong, what I ought to do were utterly disconnected from what I actually do—then it is hard to see what the point of making these judgments would be. Morality serves a purpose only if it is possible for moral judgments to have some grip on us.

It's easiest to explain this point with an analogy. Let's consider a (partially) hypothetical story about the rise and fall of the norms of etiquette. Suppose there were a group that conformed to a rigid set of rules of etiquette for dinner parties. These rules proscribed eating one's salad with the entrée fork, placing one's glass on the right side of the place setting, and so forth. Everyone in the community recognizes these rules as valid. But over time agents begin to bother less about them; conformity to them begins to drop off, people talk about them less, and they come to play an altogether less pervasive role in the community's dinners. When asked, everyone in this community can cite the relevant rules perfectly well; it's just that, as we might put it, they care less about the rules.

Then along comes a philosopher of etiquette, who is terribly concerned to show these agents that they should return to their earlier concern with these norms. How would the philosopher motivate these agents to preserve these norms? Well, it wouldn't be enough to state that the rules of etiquette are such and such. After all, these agents already know what the rules are; they just aren't concerned to conform to

intending to do X at t, and her so acting" (Scanlon 2003, 12). See also Wedgwood (2007, 32), who offers the following version of motivational judgment internalism: "Necessarily, if one is rational, then, if one judges 'I ought to ϕ', one also intends to ϕ."

[35] See Brink (1986) and Svavarsdottir (1999 and 2006). For responses to these kinds of objection, see Wedgwood (2007).

them. So what is the proponent of etiquette to do? The answer seems clear: he needs to connect the rules to something that these agents care about, to some of their goals or motives or aspirations. Perhaps they take pleasure in the pomp of formal dinners; perhaps the norms promote a desired social cohesion; perhaps they aspire to preserve tradition; and so on. These kinds of considerations might restore etiquette to its former role.

I suggest that morality is analogous. If agents develop a skeptical attitude toward their moral code, we need to be able to say something to them. It won't do simply to insist that these just are the rules of morality, any more than it would do to insist that these just are the rules of etiquette; the agents are perfectly well aware of what the rules are. We need to offer some explanation of why these rules should have a grip on the agents. Absent such an explanation, it's hard to see why the rules shouldn't just wither away, in much the way that the more recherché rules of etiquette have, to a considerable extent, withered away. (Notice that I am not claiming that our moral code actually would wither away. We have many powerful motives to conform to morality: sentiments of compassion, a desire to do what is in our long-term self-interest, a desire to be an accepted member of the community, and so forth. I am claiming that *absent* such a connection between motives and morality, morality might die out.)

Regardless of whether particular normative judgments are necessarily motivating, the etiquette story shows that it is possible for entire systems of norms to lose their connection to motivation. And we want to know whether moral norms—i.e., purportedly universal norms—might meet a similar fate. To avoid that fate, it looks like morality needs to have some connection to our motives. As Harry Frankfurt puts it, it looks like "what we *should* care about depends upon what we *do* care about" (2006, 24).

These considerations suggest that moral norms have to be grounded in our motives.[36] But there is a tension: we also want morality to provide a check on our motives. Let me explain.

Nietzsche phrases a version of this objection in terms that are somewhat unfamiliar to contemporary ethicists: he calls it the problem of nihilism. I think Nietzsche's remarks on nihilism constitute a very powerful challenge for ethical theory, but to appreciate its strength the point must be put carefully. The groundwork for Nietzsche's views on nihilism emerged from earlier nineteenth-century discussions of value, so it helps to start there. Consider a passage from Hegel's *Philosophy of Right,* in which Hegel considers the view that all value is grounded merely in arbitrary choices of the individual:

[36] Williams (1981) famously argues that an agent has a reason to ϕ only if A would be motivated to ϕ if he deliberated in a procedurally rational way from his existing motives. The idea, here, is that a consideration can be a reason only if it can motivate me; moreover, a consideration can motivate me only if it bears an instrumental connection to my existing motives.

This implies that objective goodness is merely something constructed by my conviction, sustained by me alone, and that I, as lord and master, can make it come and go. As soon as I relate myself to something objective, it ceases to exist for me, and so I am poised above an immense void, conjuring up shapes and destroying them. (Hegel 1991, §140A)

Hegel here argues that if the authority of my values arose merely from my arbitrary, unconstrained acts of will, then these values would not appear as objective constraints. Rather, they would appear as empty, ephemeral shapes—for the agent could rescind the value's authority as easily as she could bestow it. Kierkegaard makes the same point in *The Sickness unto Death*. In this work, Kierkegaard mocks the idea that the authority of values could be grounded in the agent's own doings. He claims that if "the self exerts the loosening as well as the binding power"—that is, if the authority of values consists in the agent's binding herself to these values, and if the self can loosen anything it binds, then:

The self is its own master, absolutely its own master . . . On closer examination, however, it is easy to see that this absolute ruler is a king without a country, actually ruling over nothing; his position, his sovereignty, is subordinate to the dialectic that rebellion is legitimate at any moment. Ultimately it is arbitrarily based upon the self itself. Consequently, this despairing self is forever building only castles in the air . . . just when it seems on the point of having the building finished, at a whim it can dissolve the whole thing into nothing. (Kierkegaard 1983, 69–70)

Here, Kierkegaard claims that a value whose authority is grounded merely in unconstrained choices is no value at all: if "rebellion is legitimate at any moment"—in other words, if I can reject the value as soon as I feel like doing so—then the value does not constrain me, and amounts to nothing more than a whim.[37]

I think we can fairly summarize Hegel and Kierkegaard's point as follows: in order for us genuinely to will something, in order for our goals to inspire real allegiance, we need to see these goals as having more authority than mere whims. We need to see something as non-arbitrarily structuring and constraining our choices. (What counts as arbitrary structuring is going to vary: the voice of tradition and authority used to be enough for us, but, Nietzsche thinks, no longer is.)

Nietzsche's discussions of nihilism build on this point. He offers the following definition of nihilism: "What does nihilism mean? That the highest values devalue themselves. The goal is lacking; 'why?' finds no answer" (KSA 12:9[35]/WLN 146). In other words, the values that were formerly regarded as highest or most central are experienced as unsupportable. Nietzsche explains that "this realization is a consequence of the cultivation of truthfulness—thus itself a consequence of the faith in morality" (KSA 12:10[192]/WLN 205). That is, values devalue themselves in the

[37] Hegel and Kierkegaard actually have a deeper target: the Kantian account of normativity. They argue that the Kantian account, according to which normative authority issues from volition, provides no substantive constraints; to put the point briefly, Kant's categorical imperative is an "empty formalism." Accordingly, the complaint above is directed at Kant: his theory makes it impossible to distinguish norm and whim. I discuss this issue in more detail in Katsafanas, "The Problem of Normative Authority in Kant, Hegel, and Nietzsche."

sense that our moral code prizes truthfulness, and when faithfully pursued the commitment to truth leads us to doubt whether any of our valuations—truthfulness included—can be justified. (Mere appeals to tradition, for example, are no longer accepted.) Nihilism is the belief that no values can be justified. So, a nihilist could be someone who accepts the epistemological and metaphysical challenges discussed above, and thinks that no ethical theory can answer them.

The consequences of nihilism are far-reaching. Nietzsche describes the nihilist as holding that "life is no longer worthwhile, all is the same, all is in vain" (Z IV.11). We might put the point more clearly: *because* "all is the same"—because no values enjoy any support—life is no longer worthwhile and all is in vain. Projects, commitments, and ways of life appear unsupported, arbitrary: any way of life, any choice, any action is as good as any other.

For this reason, Nietzsche is not interested in the typical bogeymen from contemporary ethical theory, the egoist, the amoralist, and their ilk. The egoist can't see any reason to conform to morality unless doing so is in his self-interest (where self-interest is assumed to be something that can be unproblematically specified). The amoralist is fully cognizant of the accepted moral code, but isn't moved by it. These characters express the following point of view: universal normative demands would be so *hard*, so *constraining*; we need to show why we'd be motivated to live under them, instead of throwing off their yoke and enjoying freedom from them.[38]

But Nietzsche isn't worried about whether morality is hard. In fact, he's interested in something like the reverse of this position.[39] He argues that in order for us to view our actions, projects, and indeed our lives as meaningful—in order for our goals to inspire real allegiance, real sacrifice, real direction for the will—we must take certain values as authoritative.[40] From this perspective, a coherent egoist or amoralist wouldn't experience relief and freedom; the 'freedom' from universal normative demands would instead bring despair and senselessness. As Nietzsche puts it in a passage that is worth quoting at length:

[38] For an insightful discussion of the problems with this point of view, see Bergmann (1994). Note also that certain accounts of rational egoism treat the rational egoist as subject to universal normative demands. See, for example, Sidgwick (1981).

[39] Karl Jaspers remarks that "Nietzsche attacks morality in every contemporary form in which he finds it, not in order to remove men's chains, but rather to force men, under a heavier burden, to attain a higher rank" (Jaspers 1997, 140). Nietzsche writes, "Basically I abhor every morality that says: 'Do not do this! Renounce!' . . . But I am well disposed towards those moralities that impel me to do something again and again from morning to evening, and to dream of it at night, and to think of nothing else than doing this *well* . . . " (GS 304).

[40] I mean this to be an uncontroversial point about Nietzsche, so let me distinguish two claims: (1) in order to avoid nihilism, we must treat certain values as authoritative; (2) treating a value as authoritative involves or requires viewing the valuation as justified. Nietzsche clearly holds (1), as the quotations below indicate. Whether he holds (2) is more controversial. We might, for example, read Nietzsche as attempting to affirm certain values without thinking that this affirmation can be justified. In Chapter 6, I will argue that Nietzsche does indeed hold (2).

What is essential 'in heaven and on earth' seems to be, to say it once more, that there should be *obedience* over a long period of time and in a *single* direction: given that, something always develops, and has developed, for whose sake it is worth while to live on earth; for example, virtue, art, music, dance, reason, spirituality—something transfiguring, subtle, mad, and divine. The long unfreedom of the spirit, the mistrustful constraint in the communicability of thoughts, the discipline thinkers imposed on themselves to think within the directions laid down by a church or court, or under Aristotelian presuppositions, the long spiritual will to interpret all events under a Christian schema and to rediscover and justify the Christian god in every accident—all this, however forced, capricious, hard, gruesome, and anti-rational, has shown itself to be the means through which the European spirit has been trained to strength, ruthless curiosity, and subtle mobility, though admittedly in the process an irreplaceable amount of strength and spirit had to be crushed, stifled, and ruined (for here, as everywhere, 'nature' manifests herself as she is, in all her prodigal and indifferent magnificence which is outrageous but noble) . . . Slavery is, as it seems, both in the cruder and in the more subtle sense, the indispensable means of spiritual discipline and cultivation, too. Consider any morality with this in mind: what there is in it of 'nature' teaches hatred of the *laisser aller*, of any all-too-great freedom, and implants the need for limited horizons and the nearest tasks—teaching the *narrowing of our perspective*, and thus in a certain sense stupidity, as a condition of life and growth. You shall obey—someone and for a long time: *else* you will perish and lose the last respect for yourself—this appears to me to be the moral imperative of nature which, to be sure, is neither 'categorical' as the old Kant would have it (hence the 'else') nor addressed to the individual (what do individuals matter to her?), but to peoples, races, ages, classes—but above all to the whole human animal, to *man*. (BGE 188)[41]

So Nietzsche is mounting a practical challenge for morality, but not the one expressed by MJI; Nietzsche does not care whether particular normative judgments are necessarily motivating. What concerns him is whether the whole system of normative judgments might become detached from our practical deliberations. What concerns him is whether normative judgments might come to seem as nothing more than the expressions of mere whims. To put the point in terms of my etiquette story from above: Nietzsche is worried that all purportedly universal normative claims might come to seem vestigial, like the norms of etiquette in our hypothetical community. If so, Nietzsche thinks, the results would be disastrous: if we become incapable of seeing any value as authoritative, if all valuations are seen as nothing more than expressions of contingent whims, then we will lose any ability to sustain our commitment to goals. As Nietzsche puts it in the quotation cited earlier, nihilism means that "The goal is lacking; 'why?' finds no answer" (KSA 12:9[35]/WLN 146). Goals must "inspire faith" (KSA 12:9[35]), but the nihilist cannot see any reason for this faith: "'Why did we ever pursue any way at all? It is all the same.' Their ears appreciate the preaching,

[41] Compare the following remark: "it is almost always a symptom of what is lacking in himself when a thinker senses in every causal connection and psychological necessity something of constraint, need, compulsion to obey, pressure, and unfreedom; it is suspicious to have such feelings—the person betrays himself" (BGE 21).

'Nothing is worthwhile! You shall not will!'" (Z III.12; cf. GS 1, GS 125, GM III.28).[42]

In short: we need something that structures our actions, categorizing certain goals as more important than others, some as worth pursuing, and so forth. Morality aspires to be just this. So there's an odd demand: to grip us, we want to say, morality must have some connection to our motives, goals, and aspirations. But the connection can't be *too* tight: for we want morality to provide some kind of check on our motives, goals, and aspirations. Or, put differently: we want morality to be related to what we care about, but we also want it to provide constraints on what we care about. A successful ethical theory must answer this challenge, by showing how normative claims can attach to our motives without collapsing into expressions of mere whim. Call this the practical challenge.

2. Assessing ethical theories in light of these challenges

So we have three challenges for moral philosophy. An adequate account of morality must answer them. In the following sections, I will consider how four dominant ethical theories fare with respect to these challenges. My claims here are not meant to be decisive; far from it. These issues have been hotly debated over the past decades, and I do not hope to resolve them in a single chapter. I intend rather to survey some familiar problems with certain popular ethical theories. I will argue that appreciating these challenges opens us to the possibility of a new kind of ethical theory, which would avoid them.

2.1 Non-reductive realism

Let's begin by considering non-reductive moral realism. Non-reductive moral realism was defended by Plato (on some interpretations), Ross (1930), Moore (1971), and Sidgwick (1981), among others. According to this view, there are moral facts and these facts are irreducible. Over the past decade or so, non-reductive moral realism has enjoyed a striking resurgence. Parfit defends this view, writing that "there are some irreducibly normative truths" (Parfit 2011, vol. II, 464). Scanlon argues for another version: he maintains that a consideration is a reason iff it "counts in favor of" some action, and he begins his book by stating that he will take this favoring relationship as an irreducible primitive (Scanlon 2000, 17).

[42] Making a related point, Pippin writes that Nietzsche is pervasively concerned with the conditions under which our "eros, our orienting commitment" is "sustainable and how it could . . . come to fail" (Pippin 2010, 11). Putting this in terms of note 40, Pippin is interested in the extent to which (1) is possible without (2); he explores the way in which we might treat values as authoritative without linking this authority to questions of justification. For a gripping discussion of this problem in a concrete context, see Lear (2006). Lear writes, "a crucial aspect of psychological health depends on the internalization of vibrant ideals . . . in relation to which one can strive to live a rewarding life. Without such ideals, it is difficult to see what there is to live for" (2006, 140). He documents the collapse and rebirth of ideals in the Crow culture.

Non-reductive realist views claim that certain normative beliefs are true. But how do we *justify* these claims about moral truths? The most common approach taken by realists is *intuitionism*, which is the view that some moral truths are knowable a priori.[43] For example, Sidgwick claims that certain moral truths are self-evident:

> the propositions, "I ought not to prefer a present lesser good to a future greater good," and "I ought not to prefer my own lesser good to a future greater good of another," do present themselves as self-evident; as much (e.g.) as the mathematical axiom that "if equals be added to equals the wholes are equal." (Sidgwick 1981, 383)

W. D. Ross agrees: "in ethics we have certain crystal-clear intuitions from which we build up all that we can know about . . . the nature of duty" (Ross 1939, 144). For Ross, these intuitions lead us to see that there are five distinct duties: fidelity, reparation for previous wrongs, gratitude, promotion of the aggregate good, and non-maleficence (Ross 1930, 19–25). More recently, Shafer-Landau has argued that we have a priori knowledge that infliction of pain on innocent children is wrong (cf. Shafer-Landau 2003), Parfit has made similar claims about the badness of suffering (cf. Parfit 2011, vol. II, 76–82), and so on.

How do these views fare with respect to the epistemological, metaphysical, and practical challenges? I contend that they do not offer convincing responses to any of these challenges. Let's start with the epistemological challenge.[44] As the above quotations demonstrate, when asked to explain why a particular normative claim is true, realists appeal to truths that are allegedly known a priori. (In a moment I will consider a complication: some realists avoid this commitment by appealing to reflective equilibrium.) But is this warranted?

On this point, Nietzschean critiques seem to me devastating. For Nietzschean genealogies should make us exceedingly skeptical of the intuitions and convictions that are being labeled "a priori knowledge." Nietzsche writes that "whoever ventures to answer" philosophical questions "by an appeal to a sort of *intuitive* perception, like the person who says, 'I think, and know that this, at least, is true, actual, and certain'—will encounter a smile and two question marks from a philosopher nowadays. 'Sir,' the philosopher will perhaps give him to understand, 'it is improbable that you are not mistaken; but why insist on the truth?'" (BGE 16). His point, here, is that genealogies

[43] Old-fashioned moral realists, such as Reid, appealed to a "moral faculty": he claimed "that by an original power of the mind, which we call conscience, or the moral faculty, we have the conceptions of right and wrong in human conduct, of merit and demerit, of duty and moral obligation, and our other moral conceptions; and that, by the same faculty, we perceive some things in human conduct to be right, and others to be wrong; that the first principles of morals are the dictates of this faculty; and that we have some reason to rely upon those dictates, as upon the determinations of our senses, or of our other natural faculties" (Reid 1983, 237). More recently, non-reductive realists have preferred to divorce claims about a priori knowledge from claims about special faculties.

[44] There is a well-known challenge for non-reductive realism at this point: we can wonder how we have epistemic access to these normative facts (cf. Mackie 1977). I will be pressing a different objection.

should debunk our confidence in our own moral beliefs, be these beliefs about suffering, equality, or what have you.

After all, our intuitions and convictions about morality are strongly influenced by the moral code under which we have been raised. Those raised in a wealthy, safe, democratic society that prizes Judeo-Christian values will have the kinds of intuitions mentioned above; surely Shafer-Landau is right to say that many of us will claim that we have a priori knowledge that killing innocent individuals is wrong. But consider Robert Pippin's objection to a related view. Discussing Mark Hauser's claim that "do as you would be done by; care for children and the weak; don't kill; avoid adultery and incest; don't cheat, steal or lie" are "moral universals," Pippin writes that this list "really takes one's breath away." He continues,

> This banal list of modern, Christian humanist values was written by a Harvard professor in a contemporary world still plagued by children sold into slavery by parents who take themselves to be entitled to do so, by the acceptability of burning to death childless wives, by guilt-free spousal abuse, by the morally required murder of sisters and daughters who have been raped, by "morally" sanctioned ethnic cleansing undertaken by those who take themselves to be entitled to do so, and one could go on and on. (Pippin 2009, 41–2)

In short, Pippin's point is that these alleged moral truths are a product of acculturation, and consequently are not universally shared. The examples discussed above illustrate this: someone immersed in the moral code of Homeric Greece would have had strong intuitions that megalopsychia, envy, social hierarchy, and so forth are good.

If we look at the way in which moral beliefs have shifted over time, if we appreciate the complex links between these concepts, it is hard to see the lists of allegedly a priori moral truths drawn up by these theorists as anything more than reports of life from within a particular moral code. The realists may well chart, systematize, and harmonize our conventional moral beliefs. But we want more than that. We want, if possible, a reason to maintain our commitment to this system.

Thus, to someone with Nietzschean sympathies—or, to put the point more polemically, to someone with historical sensitivity—the non-reductive realist simply shows us what follows from within a particular moral code. Scanlon's version of non-reductive realism, for example, shows us what follows if we take for granted the idea that we should act on principles that no one can reasonably reject; utilitarians show us what follows if we take for granted the idea that we should maximize aggregate utility. There is nothing wrong with this; it is a monumentally difficult task. But it does not so much as touch on the epistemological problem: it gives us no reason for confidence in our moral beliefs.

There is, however, a complication. So far, I have been objecting to the *intuitionist* component of non-reductive realism. This is the claim that we have a priori knowledge of certain moral truths. However, some realists eschew talk of a priori intuitions and

instead appeal to reflective equilibrium.[45] Rather than seeking to ground morality in intuitions, we might pursue a more modest task: we might try to move from the moral beliefs that we do hold to the moral beliefs that we should hold. On this view, ethical theory is a matter of increasing the degree of coherence between our beliefs about particular cases, general principles, and theoretical beliefs. We start with intuitions about which particular actions are wrong (harming innocents, lying, cheating, and so forth), which general principles are valid (happiness is to be maximized, harming is worse than failing to help, etc.), and which theoretical beliefs are true (utilitarianism is an adequate moral theory, etc.). We strive to increase the coherence among these beliefs, eliminating inconsistencies and tensions; ideally, we might even come to see some of these beliefs as providing warrant for others.

Reflective equilibrium views differ from intuitionist views in that the former need not take any particular beliefs as having a privileged epistemic status. Any belief or intuition can come up for review and possible rejection. Unfortunately, though, the epistemological problem arises in just the same way. Reflective equilibrium views need to treat some set of intuitions and beliefs as having initial credibility. After all, depending on the starting points we might end up in quite different places: there are multiple, mutually inconsistent systems of beliefs that are in reflective equilibrium. Put differently: reflective equilibrium, if faithfully executed, wouldn't lead to a unique moral code. Nietzsche's ancient nobles would have been in reflective equilibrium with a set of values that differs entirely from our own. But, given the way in which they arose, why should we grant these initial intuitions any credibility?[46,47]

So I take it that non-reductive realism fails the epistemological challenge: regardless of whether it relies on intuitionism or reflective equilibrium, it gives us no reason for confidence in our moral code.

Let's now turn to the metaphysical challenge. On the face of things, non-reductive realism is again in trouble. Non-reductive realism posits irreducible normative truths, which seem to be paradigms of properties that those with naturalistic sympathies will

[45] In his 2009 Locke Lectures (entitled "Being Realistic About Reasons"), Scanlon argues that we must employ the method of reflective equilibrium in order to defend claims about reasons for action.

[46] Some utilitarians have pressed this line: Singer criticizes Rawls for "assum[ing] that our moral intuitions are some kind of data from which we can learn what we ought to do" (Singer 2005, 346). Singer objects: "A normative moral theory is an attempt to answer the question 'What ought we to do?' It is perfectly possible to answer this question by saying: 'Ignore all our ordinary moral judgments, and do what will produce the best consequences'" (Singer 2005, 345–6). Thus, he writes that "there is no point in trying to find moral principles that justify the differing intuitions to which various cases give rise" (2005, 348). For, "there is little point in constructing a moral theory designed to match considered moral judgments that themselves stem from our evolved responses to the situations in which we and our ancestors lived during the period of our evolution . . . " (2005, 348).

[47] In *Political Liberalism,* Rawls claims that reflective equilibrium is a search for reasonable grounds of reaching agreement that can be based on "our conception of ourselves" and in our "relation to society" (Rawls 1993). But, Nietzsche would point out, our conceptions of ourselves and our relation to society have histories, and themselves embody normative claims.

find fantastical. Indeed, contemporary realists often acknowledge the counterintuitive nature of their proposal. For example, Derek Parfit writes,

> Many . . . writers ignore the possibility that there might be normative truths . . . Gibbard regards this possibility as too fantastic to be worth considering. There are good reasons to have this attitude. Irreducible normative truths, if there are any, are most unusual. As many writers claim, it is not obvious how such truths fit into a scientific world-view. They are not empirically testable, or explicable by natural laws. Nor does there seem to be anything for such truths to be about. What can the property of badness be? Given these points, it is natural to doubt whether these alleged truths even make sense. If such truths are not empirical, or about features of the natural world, how do we ever come to understand them? If words like 'reason' and 'ought' neither refer to natural features, nor express our attitudes, what can they possibly mean? (Parfit 2006, 330)

Although Parfit champions a view according to which there are irreducible normative truths about what there is reason to do, he admits that these normative truths may seem "too fantastic to be worth considering." I think this is the correct attitude: we should appeal to irreducible normative truths only if we are driven to that position by the failure of other ethical theories.[48]

Let's end by considering the practical problem. It is very hard to see why we should care about these alleged normative truths. If non-reductive realism is true, then moral facts could be completely disconnected from our motives, goals, and aspirations. But how could these moral facts be of any relevance to us? They look exactly analogous to the etiquette facts discussed above: for those on whom they have no grip, it is hard to see what to say. All the realist can say is "It's a fact that ϕ-ing is wrong" or "You should do what's in accordance with objective morality."[49] We'd like a view that can do more than this, by showing why moral requirements are something we should care about.[50]

Realists do have a response. For example, Parfit argues that this kind of objection conflates normative authority with motivational force: "many people, I believe mistakenly, regard normative force as some kind of motivational force" (Parfit 1997, 126).

[48] Some realists have attempted to defuse these metaphysical worries. The "partners in guilt" argument has been a favorite: realists argue that normative properties are no more mysterious than mathematical properties. For example, Scanlon claims that moral beliefs, like mathematical beliefs, "do *not* make claims about things that exist in space and time and the causal relations between them" (2003, 9). Consequently, he reasons, moral beliefs are not in conflict with science, which Scanlon understands as an "account of the occurrence of events in the spatio-temporal world and of the causal relations between them" (2003, 9). Science and morality are about different things. Parfit offers a similar argument, writing that "there are some claims that are irreducibly normative . . . and are in the strongest sense true. But these truths have no ontological implications. For such claims to be true, these reason-involving properties need not exist either as natural properties in the spatio-temporal world, or in some non-spatio-temporal part of reality" (2011, II, 486). Other realists try to defuse the metaphysical problem in a different way: they argue that normative properties, though irreducible, are "realized" by natural properties (see, for example, Wedgwood 2007).

[49] Thus, when pressed on these kinds of questions, a realist like Clarke can do nothing more than resort to an ad hominem: "These things are so notoriously plain and self-evident, that nothing but the extremest stupidity of mind, corruption of manners, or perverseness of spirit, can possibly make any man entertain the least doubt of them" (Clarke, Boyle Lectures of 1705, reprinted in Schneewind 2002, 296).

[50] For similar complaints, see Korsgaard (1996b) and Gibbard (2003, 152–8).

Parfit's idea is that normative facts needn't have any motivational force whatsoever: "even when [normative claims] do not have motivational force, they could . . . have *normative* force" (1997, 111–12). Scanlon takes the same approach. He writes, "Suppose a person believes that he has conclusive reason to do *X* at *t*. How can this fall short of what is required?" (Scanlon 2003, 14). Scanlon grants that the person might not be moved to *X*. But he does not see this as an objection to the view. I do not find these responses convincing: they amount to a kind of mysterianism about normative authority. The response consists merely in the assertion that there is a *sui generis* property, normative authority, that can obtain independently of any facts about motivation. But this is precisely what the practical argument calls into question: in light of the historical considerations adduced above, the mere insistence that there is a *sui generis* property of normative force seems, as Nietzsche might put it, rather quaint.

2.2 Aristotelian theories

I have suggested that non-reductive realism has trouble mustering convincing responses to the three challenges. So let's turn to a different kind of ethical theory: Aristotelianism. Aristotelians argue that we can derive norms from facts about the natures of things. It's easiest to see how by considering objects that have functions—motors, toasters, knives, hearts, lungs. For any type with a function, we can evaluate particular tokens of that type with respect to whether they have the properties required to fulfill the function. A good knife is one that has the properties necessary for cutting; a bad knife is one that lacks some or all of these properties. The same goes for parts of living creatures: a heart is defective if it lacks the properties required for circulating blood in the requisite way. The central Aristotelian idea is that we can extend this kind of evaluation to living things, including human beings.

Rosalind Hursthouse gives a nice summary of this approach:

> Living things can be . . . evaluated according to all sorts of criteria. We may evaluate them as potential food, as entries in competitive shows, evens as 'decorative objects for my windowsill given my preferences,' and each noun or noun phrase brings its own criteria of goodness with it. In the context of naturalism [i.e., Aristotelianism] we focus on evaluations of individual living things as or *qua* specimens of their natural kind, as some well-informed gardeners do with respect to plants and ethologists do with respect to animals. (Hursthouse 1999, 197)

In short, these theorists offer a characterization of "natural kinds," and then evaluate particular individuals by determining whether "this individual *x* is a good *x*, a good specimen of its kind" (Hursthouse 1999, 203). Thus, this view "hopes to validate" ethical claims "by appeal to human nature" (Hursthouse 1999, 193).

There are variations in the details of these views. Hursthouse argues that we can evaluate aspects of an individual with respect to how well they contribute to four criteria: the individual's survival, the continuation of the species, the individual's pleasure and freedom from pain, and (with social animals) the functioning of the group (Hursthouse 1999, 200–3). Bloomfield (2001) contends that moral goodness is

the state of character that disposes human beings to flourish, where flourishing is defined in terms of biologically determined human purposes. As he puts it, "moral properties have the same ontological status as healthiness or other biological properties" (Bloomfield 2001, 28). Thomson (2008) relies on the idea that there are "goodness-fixing kinds": kinds such that being a member of that kind establishes standards of excellence for its members.

How do these views fare with respect to the epistemological, metaphysical, and practical challenges? They have a good response to the epistemological challenge: unlike the non-reductive realist, the Aristotelian need not place any great faith in intuitions, conventional beliefs about morality, and so forth. She can simply appeal to natural kinds or biologically defined functions. If these functions or kinds can be specified in a way that does not presuppose the truth of particular moral claims, then the Aristotelian has a way of stepping outside of and assessing her current moral beliefs. Provided the critical reflection supports the moral code, the epistemological challenge will have been answered.[51]

This brings us to the metaphysical challenge: is the Aristotelian view consistent with naturalistic strictures? This depends on the particular version of the Aristotelian theory that we embrace. A burden for the Aristotelian is to show why we should believe that *human being* is a normative kind. Aristotle himself bases his argument on claims about natural teleology that have been discredited by modern science (cf. Williams 1986). But we can develop a naturalistically respectable version of Aristotelianism. For example, we could appeal to the tendencies or functions instilled by biology or by natural selection (cf. Bloomfield 2001). Or we could simply note that there's nothing inherently problematic about specifying what it is for a tomato plant or a tiger to flourish; by extension, the same point should apply to human beings.[52] So the Aristotelian may be able to answer the metaphysical challenge.

The practical challenge is more difficult. Suppose I accept that human beings have a function, or that "human being" is a normative kind. Why should this matter to me? Why should I care whether I am a defective instance of my kind? After all, it seems obvious that I regularly neglect features that are characteristic of the human kind. It is probable that natural selection has instilled in me a disposition to reproduce as often as possible, to be distrustful of and somewhat hostile toward those who are not members of my immediate group, to eat as much sugary food as possible, and so on. Why should I try to realize these aims? Many of the conditions that made these dispositions beneficial in the evolutionary past are no longer present. (A disposition to eat as

[51] Some Aristotelian views start with a moralized notion of flourishing, and define the human function in terms of it. Annas (1995) and McDowell (2001, Chapter 1) adopt this strategy. These views do not have the epistemic advantage mentioned above.

[52] As Hursthouse puts it, "we evaluate ourselves as a natural kind, a species which is part of the natural biological order of things, not as creatures with an immortal soul or 'beings' who are persons or rational agents" (1999, 226). Foot writes "I am therefore . . . likening the basis of moral evaluation to that of the evaluation of behavior in animals" (Foot 2003, 16).

much sugary food as possible made sense when that aim could be realized only by eating fruits; it does not make sense when I can realize it by eating cookies.) Even when the aim hasn't had its conditions of realization altered, I can decline to fulfill it. To be sure, reproduction is fitness enhancing; but what's that to me?[53] Absent a convincing answer to this question, the Aristotelian views fail to answer the practical challenge.[54]

2.3 Humean theories

I have suggested that non-reductive realism has trouble with the epistemological, metaphysical, and practical challenges. Aristotelianism has the potential to avoid the epistemological and metaphysical challenges, but lacks a convincing response to the practical challenge. In light of this, let's consider a third view: Humeanism. Taking off from Hume's claim that nothing is "in itself valuable or despicable," apart from the "particular constitution and fabric of human sentiment and affection" (Hume 1987, 162), Humean views claim that normative facts must be explained by some conative state of the agent:

Agent A has reason to ϕ iff A has a conative state of type T that is suitably connected to ϕ-ing.

This is just a schema; it needs to be filled in by specifying what kinds of conative states are at issue and what counts as a suitable connection. Different specifications of T and "suitable connection" will generate different versions of Humeanism.

The simplest version of Humeanism claims that all actual desires provide reasons: A has a reason to ϕ iff A has a desire whose fulfillment would be promoted by ϕ-ing. This version is probably too simple: for example, it entails that desires based on false beliefs generate reasons.[55] So Humeans typically adopt a modified view. For example, Bernard Williams argues that A has reason to ϕ iff A has a conative state that is connected by a "sound deliberative route" to ϕ-ing. This "sound deliberative route" can include the correction of false beliefs, the appreciation of instrumental connections between one's desires, and so forth. On this view, we are no longer forced to say that

[53] As Copp and Sobel put it, "Why should the constituents of natural goodness for members of my species (or 'life form') determine what counts as morally good for me?" (Copp and Sobel 2004, 542). Indeed, the idea that something ought to meet the standards of its normative kind seems most plausible when there is a clear *end* that has been adopted. That is, we will agree that this object is a defective toaster because we want this object to serve a particular end: making toasted bread. But if an agent didn't have that end—if I merely had the end of collecting shiny metal objects—then the fact that this toaster doesn't heat up bread would hardly seem to count as a defect. Put differently, we might argue that talk of functional kinds or normative kinds is really best explained as talk of *items that are presumed to serve a certain end*. Absent commitment to that end, it is hard to see why we should care about them.

[54] Proponents of this view typically respond in just the same way as non-reductive realists: they insist that normative authority need not translate into motivational force. See Foot (2001) and Thomson (2008). This inherits exactly the same difficulties as the non-reductive realists' response, discussed at the end of Section 2.1.

[55] Take Bernard Williams' classic example (Williams 1981, 101–13): I desire to drink that cup of liquid, which I believe is full of gin. However, the cup has actually been filled with gasoline. If I knew that the cup was full of gasoline, I wouldn't desire to drink it. In this case, it hardly seems that I have a reason to drink the liquid. In general, desires that are based upon false beliefs do not seem to provide reasons.

desires based on false beliefs generate reasons.[56] And once we set off on this track—moving from facts about the agent's actual desire to facts about what the agent would desire in suitably different circumstances—we might be tempted to go even further. For example, Michael Smith argues that moral facts are facts about what a perfectly rational and fully informed version of my actual self would desire that my actual self do (Smith 1994).[57]

Let's consider how Humean views fare with respect to the three challenges. The metaphysical challenge presents no difficulties: Humeans need not appeal to queer normative properties, but only to desires and other motivational states. There is nothing queer or non-natural about the idea that agents have various desires, affects, and motivational states. Moreover, Humeans have a straightforward response to the practical challenge: if we analyze normative claims in terms of the agent's motivational states, it's easy to see why agents would be motivated by normative claims. (Of course, the connection between motivation and reasons will become more tenuous to the extent that—like Michael Smith—we analyze reasons in terms of merely hypothetical or idealized motives.)

However, things go less smoothly when we consider the epistemological point. Some Humeans present their theories as immune to epistemological worries. For example, Hume himself remarks that although skeptical arguments "may flourish and triumph in the schools," in ordinary life "they vanish like smoke, and leave the most determined skeptic in the same condition as other mortals" (*Enquiry*, 159). The skeptic, he tells us, "cannot expect, that his philosophy will have any constant influence on the mind . . . Nature is always too strong for principle" (*Enquiry*, 160).[58] But there are reasons to doubt that the Humean theories are secure in light of the Nietzschean epistemological point. We want to know whether we have a reason for confidence in our moral code. Above, we saw that intuitions and moral beliefs provide poor grounds for confidence. An analogous problem arises for conative states. Nietzsche writes:

> *Feelings and their origin in judgments*.—'Trust your feelings!'—But feelings are nothing final or original; behind feelings there stand judgments and evaluations which we inherit in the form of feelings (inclinations, aversions). The inspiration born of a feeling is the grandchild of a judgment—and often a false judgment!—and in any event not a child of your own! To trust one's feelings—means to give more obedience to one's grandfather and grandmother and their grandparents than to the gods which are in *us*: our reason and our experience. (D 35)

[56] Williams (1981, Chapter 8).

[57] For a sophisticated defense of a Humean view, see Schroeder (2007).

[58] Charles Griswold notes that analogous points apply to Adam Smith: "Smith does not hold that as moral actors we normally treat morality as a skeptic would. Rather, we act as though commonsense moral realism were valid, that is, as though moral qualities exist objectively in the nature of things, are external to us and claim authority over us" (Griswold 1999, 165). As Griswold puts it, for Smith and Hume "skeptical theorizing ought not to budge our everyday beliefs" (1999, 165).

Here, Nietzsche claims that feelings often originate in judgments.[59] Let's illustrate this with an example. Marrying one's first cousin was quite common in the ancient world, and is still widely practiced in certain parts of the Middle East and Sub-Saharan Africa. However, most individuals in the United States and Europe view cousin-marriage as disgusting or even repellent. We tend to concoct justifications for this emotion. For example, we tell ourselves that cousins who marry are more likely to have children with birth defects. However, this is demonstrably false; cousins are no more likely to have genetically defective children than non-cousins.[60] I suspect that even upon appreciating the falsity of this belief, most individuals in Western societies will continue to view cousin-marriage as disturbing or even disgusting. This provides an example of the way in which an evaluative belief—that marrying one's cousin is wrong—gradually generates a variety of affects (disgust, revulsion, etc.), which are resistant to transformation, persisting even in the absence of evidence for the belief.[61]

If our conative states are influenced by our evaluative judgments, then we cannot treat them as providing any grounds for confidence in these evaluative judgments. Consider the dialectical situation: we want to know why we should keep our promises, be compassionate, value equality, and so forth. The Humean says: abide by these rules because doing so accords with your conative states. And the Nietzschean responds: it's true that our conative states and avowed normative claims will tend to be in general conformity with one another, but this is simply due to the fact that they are reciprocally influencing. If we lack confidence in the normative claims, we should also lack confidence in the conative states.[62]

In this respect, the Humean and non-reductive realist projects can be seen as mirror images of one another: the Humean project of constructing morality through an examination of the conative states won't inspire confidence in our normative claims, given that we can modify the conative states, and a non-reductive realist project of

[59] The same point is made in the following passage: "whence come evaluations? Is their basis a firm norm, 'pleasant' or 'painful'? But in countless cases we first *make* a thing painful by investing it with an evaluation" (KSA 10:24[15]).

[60] For a helpful discussion of this case, see Prinz (2007, 240). The genetic data are given by Bennett et al. (2002).

[61] The reader is invited to try this experiment in a class: ask students whether marrying one's first cousin is wrong. Students almost inevitably say that it is. When asked *why* cousin-marriage is wrong, students typically respond by citing the alleged potential for genetic defects. When told that this belief is false, students tend not to revise their moral judgment. Instead, they resort to saying that cousin-marriage is revolting or disturbing. Here we have exactly the process that Nietzsche describes: a moral evaluation—based on any superstition, custom, or false belief—generates a strong affect; the affect is then taken to justify the moral evaluation that caused it. For a classic discussion of this phenomenon, see Haidt (2001).

[62] In fact, things are even worse. Nietzsche posits an additional explanatory factor: he maintains that it is not only judgments which influence our conative states, but a drive to imitate our fellows. He writes, "It is clear that moral feelings are transmitted in this way: children observe in adults inclinations for and aversions to certain actions and, as born apes, *imitate* these inclinations and aversions; in later life they find themselves full of these acquired and well-exercised affects and consider it only decent to try to account for and justify them . . . " (D 34). Nietzsche calls this the "herd instinct": he believes that individuals have a strong drive toward conformity and imitation. For an illuminating discussion, see Richardson (2004, 81–95). If Nietzsche is correct about the herd instinct, we have even less ground for confidence in our conative states.

justifying morality through intuition of irreducible normative truths won't inspire confidence, given that we can modify our intuitions.

Indeed, Nietzsche speculates that appreciation of this fact will gradually transform our normative judgments and conative states. Nietzsche writes:

> We have to *learn to think differently*—in order at last, perhaps very late on, to attain even more: *to feel differently*. (D 103)

Just as our moral beliefs and intuitions are nothing "final and original," nothing worthy of trust or confidence, so too our conative states must be called into question. The Nietzschean critique reveals that our beliefs and conative states are deeply intertwined and subject to the vicissitudes of history. So neither beliefs nor conative states provide a stopping-point for questions about justification. Showing that some evaluation is consistent with our conative states does not show that we have any reason to embrace the valuation; showing that some conative state is consistent with our evaluations does not show that we have any reason to embrace it.

Certain Humeans will accept this argument and claim that it only shows that there are no universal reasons. This is one possible response. It amounts to a denial that we have any grounds for confidence in our current moral system. Perhaps this is the best we can do. If so, we won't have any answer to the epistemological challenge.

Other Humeans do try to meet this challenge. For another way to avoid the worry is to show that there are some universal reasons. For example, on Michael Smith's view normative claims are contingent upon the motives of particular agents; however, he argues that universal moral claims could be true if all rational and fully informed persons would converge on a common set of desires (Smith 1994, 187–9). Below, I will argue that something like this approach can work. First, though, I want to consider one last approach to ethics.

2.4 Kantian theories

Let's end by considering Kantianism. Kant attempts to anchor universal normative claims in facts about agency. An outline of the Kantian argument would go something like this: we are committed to acting autonomously. Acting autonomously requires acting on a law or principle. The law cannot be hypothetical, i.e., tied to the realization of some goal or the satisfaction of some inclination, because the will would then be determined to action by something external to itself (i.e., an inclination or goal). Instead, the law must be categorical; it must be unconditionally valid. Kant states the content of this law as follows: "act only in accordance with that maxim through which you can at the same time will that it become a universal law" (G 4:421). He argues that this law—the Categorical Imperative—rules out certain actions, thereby yielding determinate constraints on permissible actions. So, Kant moves from a claim about agency—that we are autonomous—to a normative claim about what we have reason to do (i.e., act on laws that are in accordance with the Categorical Imperative).

This argument is notoriously difficult. Chapter 4 will examine a version of it in more detail. For now, the details of the argument won't matter: we need only consider the basic *structure* of the argument in order to see how it fares with respect to the three challenges.

If Kantianism worked, it would avoid the epistemological problem: we would start with facts about our agential nature and show that universal normative claims issue from them. As these norms follow from facts about the nature of rational agency as such, they will apply to all agents, regardless of the particular moral code that these agents currently embrace. Kantianism would thus give us a way of stepping outside of and assessing our current moral beliefs. Likewise, Kantianism has no difficulty with the practical challenge: normative claims get their grip on us because we are committed to acting autonomously, and these norms simply specify what we have to do in order to be autonomous agents. Finally, the Kantian theory might avoid the metaphysical problem as well: if we can ground normative claims in a naturalistically respectable account of agency, then there will be no need to appeal to queer metaphysical properties.

I have critiqued non-reductive realism, Aristotelianism, and Humeanism for facing certain structural problems that render them at best unlikely to answer the three challenges. Kantianism avoids this problem: it has the right *structure* to answer the various challenges. Should we then be Kantians?

I think not. For the Kantian theory faces a number of *internal* problems. (The following chapters discuss some of these problems in detail; for now, I simply mention them.) These center on its analysis of agency and its attempt to extract normative content from this analysis.

Start with the first point. Although some Kantians present their accounts of agency as naturalistically respectable, it is far from clear that all is well on this score.[63] Consider, for example, Kant's reliance on the idea that reason can "of itself, independently of anything empirical, determine the will" (*Critique of Practical Reason* 5:42); or that all actions are done on maxims (*Groundwork* 401n, 421n). Both of these claims look implausible in light of recent empirical psychology.[64] Additionally, some argue that the Kantian theory actually does require substantive and implausible metaphysical commitments, such as a claim that we have a kind of freedom consisting in "independence from everything empirical and so from nature generally" (*Critique of Practical Reason* 5:29, 97).[65] In short: problems may arise in the Kantian's reliance on a highly questionable model of agency.

Second, even brushing those problems aside, the Kantian arguments that attempt to move from the nature of agency to normative conclusions face objections at each turn: it is notoriously difficult to show how commitment to the Categorical Imperative is

[63] Korsgaard (2010) is a good example. For arguments that aspects of the Kantian enterprise can be rendered naturalistically respectable, see Scheffler (1992) and Velleman (2006).

[64] I will discuss these problems in Chapter 5.

[65] See Wood (1984, 74–83), Allison (1990, 227–9), and Schneewind (1998, Chapters 22–23).

supposed to follow from Kant's initial conception of agency, and even if we can do *that*, there are reasons for doubting that the Categorical Imperative can generate any substantive conclusions about what there is reason to do.[66] Thus, although Kantian ethics has the right *structure* to avoid the three challenges, there are reasons for doubting the cogency of the theory.

3. Constitutivism

Above, we have examined four chief competitors in ethical theory: non-reductive realism, Aristotelianism, Humeanism, and Kantianism. I briefly raised some problems for each of these theories. (Again, I do not intend the brief discussions above to constitute decisive refutations of the various theories. I intend them merely to indicate some potential difficulties that the theories would have to overcome.) I suggested that non-reductive realism runs headlong into the epistemological, metaphysical, and practical problems. Aristotelian views avoid the epistemological problem, but may encounter metaphysical difficulties and lack a convincing response to the practical problem. Humeanism avoids the metaphysical and practical problems but faces the epistemological challenge. Kantian theories avoid the epistemological and practical problems, but may be premised upon metaphysically untenable claims about agency. In sum:

	Epistemological	Metaphysical	Practical
Non-reductive realism	X	X	X
Aristotelianism		?	X
Humeanism	X		
Kantianism		?	

I think a version of this dialectic drives some philosophers toward Kantianism. After all, Kantianism faces enormous internal difficulties; it can hardly be denied that the arguments are obscure, the challenges severe. But if you think that all other ethical theories fail—if you think that in some cases they do not even *aspire* to answer the genuine puzzles—then it makes sense to embrace a problematic theory. For Kantianism at least has the virtue of confronting and attempting to answer the three challenges above. It doesn't shirk from these challenges and doesn't reduce its ambitions. If its success looks dubious, it at least sets off on the right track. So, at any rate, it seems to me.

[66] Hegel is the *locus classicus* for this objection; see his *Philosophy of Right* Section 135. In the *Phenomenology*, Hegel puts the point this way: "It would be strange, too, if tautology, the principle of contradiction, which is admitted to be only a formal principle for the cognition of theoretical truth, i.e., something which is quite indifferent to truth and falsehood, were supposed to be more than this for the cognition of practical truth" (*Phenomenology*, Section 431). Wood (1990) offers a helpful discussion of the formalism objection.

Should we then cast our lot with the Kantians? I think not. For lately a distinct kind of ethical theory has emerged. This theory, which is often called *constitutivism*, offers a fresh start. It has the advantages of Kantianism without the problems. Below, I introduce the theory and indicate how it has the potential to overcome the three challenges. I then ask how closely related constitutivism and Kantianism are.

3.1 Introducing constitutivism

To see how constitutivism works, we need to reorient ourselves. We have been examining universal normative claims that apply to action as such. But let's approach our topic from a different angle: let's consider norms pertaining to more restricted kinds of activities.

Certain kinds of activities are distinguished by the fact that participants in these activities *necessarily* have a particular aim. There are simple examples of this phenomenon, such as the game of chess. Arguably, it is not sufficient to count as playing chess that one simply moves one's chess pieces around on the board in accordance with the rules of chess. In addition, one must aim at achieving checkmate.[67] If you do not have that aim—if you are just moving pieces, without aiming to win—then you are not really playing chess. Thus, the aim of checkmate is non-optional for chess players: if you are playing chess, then you have the aim.

Of course, the aim of checkmate can be influenced and modified by other factors. But it cannot be wholly abandoned. Consider an example. If you are playing chess with a child who is just learning the game, you may also adopt the aim of letting her have a fair chance at winning. This aim will modify the way in which you pursue the aim of achieving checkmate. For example, you may see a way to achieve checkmate, but decline to take it, in order to give the child a better chance of winning. But this kind of deviation from the activity's aim can only go so far, lest you cease to engage in the activity of playing chess. If you are not making *any* effort to achieve checkmate, then you are not *really* playing chess. Instead, you are engaged in a more complex activity, with a different aim: you are engaged in the activity of *teaching a child how to play chess*, or some such. (Notice that if you are not pursuing the aim of checkmate at all, the child could justifiably complain that you are not really playing chess.)

Similarly, some philosophers have argued that the attitude of belief aims at truth.[68] For it seems that each instance of belief aims at truth, and aiming at truth is part of what constitutes an attitude as a case of belief. After all, if an attitude had absolutely no tendency to be responsive to indications of its truth value—if, for example, an attitude with the content *that p* persisted despite the agent's appreciation of conclusive evidence that *not p*—then the attitude would not be a belief.

[67] I am simplifying a bit: one could also aim at achieving a draw.

[68] For two examples, see Shah (2003) and Shah and Velleman (2005). For more skeptical discussions, see Wedgwood (2002) and Owens (2003).

Let's be more precise. We can define *constitutive aim* as follows:

(Constitutive Aim) Let A be a type of attitude or event. Let G be a goal. A constitutively aims at G iff
(i) each token of A aims at G, and
(ii) aiming at G is part of what constitutes an attitude or event as a token of A.[69]

For example, suppose we let A be the attitude of belief and G be truth. Then belief has a constitutive aim of truth iff (i) each token of belief aims at truth, and (ii) aiming at truth is part of what constitutes an attitude as a belief.

We now have an account of constitutive aims. But what would follow from the fact that chess, belief, or some other type of attitude or event has a constitutive aim? Well, suppose we accept a relatively uncontroversial claim:

(Success) If X aims at G, then G is a standard of success for X.

For example, if chess players aim at checkmate, then we can evaluate chess players with regard to whether their actions are conducive to their goal of achieving checkmate. Or, if belief aims at truth, then we can evaluate processes of belief formation in terms of how well they promote the goal of believing truths.

Note that Success simply claims that aims generate standards of success. It applies to all aims, not just constitutive aims. Whenever you have an aim, you have a standard of success.[70] Take our aforementioned chess player. Suppose she has the aim not only of checkmating her opponent, but also of enjoying her game. Then we get two standards of success: we can evaluate a particular move with regard to whether the move brings her closer to checkmate, and whether it makes the game enjoyable. These aims can interact with and modify one another: if move A would promote checkmate yet would be boring, while move B would be fascinating yet somewhat more risky, then the player may have reason to make move B. Thus, the reasons induced by the constitutive aim will be one source of reasons among many others.

So what's special about *constitutive* aims? The reasons derived from the constitutive aim differ from these other reasons in that they are *intrinsic* to the activity in question. You can play a chess game without aiming to enjoy it, and a chess game is not necessarily defective if not enjoyed. But you can't play a chess game without aiming to achieve checkmate, so a move in a chess game is necessarily defective if it does not

[69] Here it is worth making two points about the definition. First, condition (i) is implied by condition (ii). Strictly speaking, then, condition (i) is superfluous. I include (i) as a separate condition merely for the sake of clarity. Second, nothing important hinges on my restriction of A to attitudes and events; I would be happy to include other categories that might have constitutive aims. I cite attitudes and events simply because these are the categories that have been thought to possess constitutive aims. (I am including actions under the broader category of events.)

[70] In the next chapter, I will consider some objections to this claim. I will also explicate the difference between aims and desires. In Chapter 7, I show that the version of constitutivism that I defend relies only on the following principle, which is even less controversial than Success: if an agent aims at G, *and the agent endorses this aim*, then G is a standard of success for the agent's action.

promote the goal of checkmate. Thus, the interesting feature of constitutive aims is that, being inescapable, they generate *intrinsic* standards of success.

As a result, these standards readily meet challenges to their authority. If someone is engaged in an activity that these standards govern, then there is a ready answer to the question "why should I care about these standards?" The answer is just this: insofar as you are committed to this activity, you are committed to those standards. For example, a person who is playing chess has a good reason to abide by the standards constitutive of chess: if he doesn't abide by them, he will no longer be playing chess.

To see why this is important, it helps to contrast constitutive standards with other types of standards. Consider familiar rules such as "No smoking in this restaurant" or "Provide 24-hour notice of any changes to your doctor's appointment." These rules govern activities such as dining and making appointments. But one does not need to obey these rules in order to participate in the activities: I can light up in the restaurant, and I can cancel my appointment an hour in advance. To be sure, I may face penalties for failing to respect these rules. However, I do not cease to participate in the activity of eating at a restaurant simply because I light a cigarette; nor do I cease to engage in the activity of making a doctor's appointment simply because I change the appointment an hour in advance. Chess is different: if I do not govern myself with chess's constitutive standards (by trying to capture pieces, move bishops on diagonals, and so on), then I will not be playing chess at all.[71]

Suppose someone asks, "Why should I care about providing 24-hour notice when changing my doctor's appointments?" Of course, there are answers to the question—answers invoking the financial penalties that the canceled appointment will produce, the inconvenience to the doctor and her other patients, and so on. But notice that these answers invoke *external* standards. The standards apply because medical appointments are related to other activities, goals, and practices that concern the agent. The standards governing chess do not have that feature: we can answer the question "Why should I care about capturing your king?" simply by referring to the rules constitutive of the game. Thus, the chess player should care about capturing the king because if he doesn't govern himself by this standard, he won't be playing chess.

So this is the intriguing feature of the standards induced by constitutive aims: they are internal to the activity in question. Accordingly, we need not invoke external facts in order to legitimate their claim to authority.[72]

[71] One question that arises concerning constitutive rules and games is whether games have any *non-constitutive* rules. To answer this question, we would need to determine whether we could eliminate rules without thereby changing the game. For example, if we eliminated the rule stating that pawns can be moved two spaces on the first move, would participants in the resulting activity still be playing chess? I will return to this question in the next chapter.

[72] External facts will be relevant, of course. If I am engaged in a game of chess, and suddenly notice that my house is burning down around me, then there's a very real sense in which my reasons for capturing my opponent's queen are outweighed or silenced by my reason to stop playing and call the fire department. I will return to this point in Chapter 2.

Let's now take a step back. We have been considering particular types of action, such as chess-playing. But suppose we could show that *action itself* has a constitutive aim. If every agent shares a common aim, then every agent shares a common standard of success. The reasons generated by this standard will be universal: they will apply to all agents, regardless of facts about the agents' contingent desires and circumstances. Just as all chess players have a reason to checkmate, so too all agents will have a reason to fulfill the constitutive aim of action. Accordingly, the reasons generated by action's constitutive aim would have the right *form* to be moral reasons; they would be universal.[73]

So we have a two-step recipe for a new moral theory. First, we need to show that action has a constitutive aim. Second, we appeal to some version of Success in order to derive reasons from this constitutive aim. This would anchor universal reasons in facts about aims that are constitutive of agency.

Easier said than done, of course. Showing that action has a constitutive aim is going to be enormously difficult, and defending a version of Success will raise puzzles of its own. But let's set these worries aside, for a moment, and ask whether it's even worth embarking on these tasks. Would a successful version of constitutivism answer the three challenges?

3.2 Constitutivism and the three challenges

We can start with the epistemological challenge: would constitutivism give us a reason for confidence in our normative beliefs? It would. If action had a constitutive aim, then this aim would generate normative claims with a universal status. Cultural and historical variation in moral codes would not be troubling, for we would have a standard against which we could measure and critique these variations. In short, we could say that some variations are mistakes.

Moreover, notice that Success is an exceedingly spare claim. It can serve as a kind of Archimedean point in debates about ethics: we disagree about whether we have reason to be compassionate, whether happiness is more important than duty, whether suicide is wrong, and so forth. But we can set aside this disagreement on substantive ends and agree on this entirely procedural or structural conception of rationality: we can agree that if you have an end, you should strive to fulfill it, while disagreeing about what those ends are.[74]

Next, consider the practical challenge. Again, this is easily overcome: the constitutivist has no more trouble with this than does the Humean or Kantian. Norms will issue from our aims, and hence will be things that we are motivated to meet. There is no puzzle about why chess players are motivated to achieve checkmate; just

[73] It is, of course, a further question whether the universal reasons generated by the constitutive aim will be the ones we expect. Pre-theoretically, we expect the universal reasons to include claims such as "you should not murder" and "harming innocent people for fun is wrong." Whether action's constitutive aim entails these particular normative claims will depend upon what, exactly, action's constitutive aim is.

[74] I will consider objections to Success in the next chapter.

so, there would be no puzzle about why agents are motivated to achieve action's constitutive aim.

What about the metaphysical challenge? Here things are a bit more complex. Whether constitutivism meets this challenge depends on the particular version of constitutivism that we embrace. In order to provide an account of reasons that is compatible with naturalistic strictures, the constitutivist account must be grounded in a naturalistically acceptable account of agency. Suppose Nietzsche and other philosophers are correct in claiming that Kant's theory of agency is indefensible; then it would do no good to show that we can extract a constitutive aim from the Kantian conception of agency, precisely because nothing in the world answers to that conception. What we need to do, instead, is start with an accurate description of actual *human* agency.

For this reason, we should be wary of trying to demonstrate the presence of this aim via conceptual analysis. As I will explain in the next chapters, the dominant versions of constitutivism often appear to start with claims about our *concept* of agency, and to show that we can extract a constitutive aim from this concept. For example, Korsgaard writes that "it is essential to the *concept* of agency that an agent be unified," and attempts to derive normative conclusions from this alleged fact (2009, 18; emphasis added).[75] If constitutivism relies on an uncritical faith in our current conception of agency, then it won't answer the epistemic challenge. Our concept of agency is something that itself has a history: like our beliefs about morality, our intuitions about agency have undergone substantial changes over time.[76] Thus, Nietzsche criticizes thinkers who "accept concepts as a gift . . . as if they were a wonderful dowry from some sort of wonderland," rather than recognizing that they are "the inheritance from our most remote, most foolish, as well as most intelligent ancestors," and therefore stand in need of "an absolute skepticism" (KSA 11:34[195]/WLN 13). Put simply, our current concept of agency cannot be taken for granted and used as a starting point. Intuitions about agency have the same status as intuitions about morality: unless they can be independently supported, they give us no reason for confidence.

For these reasons, an adequate version of constitutivism must defend the account of agency upon which it relies. I believe that the most promising way of doing so is by relying on an empirical account of agency, rather than attempting to divine the structure of agency in an a priori or conceptual manner. I will pursue that strategy in the following chapters: I will argue that a roughly Nietzschean account of agency is not only empirically convincing, but also allows us to see that action has a constitutive aim.

[75] In the previous sentence, I said that constitutivist theories often *appear* to be starting with claims about the concept of agency. When we examine these theories in more detail, in Chapters 3 and 4, we will see that this appearance is misleading. For example, while Velleman does begin with some claims about what is essential to our concept of agency, he is concerned to this concept is "realized in the world"—that is, whether our concept of agency matches the reality (2000, 129).

[76] For helpful analyses of these changes, see for example Taylor (1992) and Williams (1993).

3.3 Constitutivism's relationship to Humeanism and Kantianism

So we can see that constitutivism has considerable potential: a successful version of the theory would overcome the three challenges to morality. In the next chapter I will ask whether constitutivism can be defended against a series of recent objections. Before proceeding to that, though, let's ask how constitutivism relates to the ethical theories surveyed above.

There is a clear parallel between constitutivism and Kantianism: both theories attempt to ground normative claims in facts about agency. Perhaps for this reason, constitutivism is often thought to be a Kantian theory. This is a mistake: in its most straightforward form, the constitutivist theory actually has more in common with Humeanism than Kantianism.

This claim may come as a surprise. Currently, the literature contains two worked-out versions of constitutivism. Christine Korsgaard's version of constitutivism is unambiguously Kantian: she seeks to show that "Kant's two imperatives of practical reason"—the Hypothetical and Categorical Imperatives—are "constitutive principles of action, principles to which we are necessarily trying to conform insofar as we are acting at all" (Korsgaard 2009, xii). David Velleman denies that we can extract Kant's Categorical Imperative from the concept of agency, but nonetheless claims that "the aim constitutive of agency can be seen to have pushed us in the direction of our moral way of life," making morality a "rational development, a form of rational progress" (Velleman 2009, 149). Velleman calls this a "Kinda Kantian strategy" (Velleman 2009, 149). In light of these remarks by Korsgaard and Velleman, we might assume that constitutivism necessarily takes a Kantian form.

But that would be a mistake. Constitutivism is based on the idea that action has a constitutive feature whose presence yields substantive normative content. Whether particular constitutivist theories take a Kantian form depends on the content of this feature (and perhaps also the way in which the theorist argues for the presence of the aim).

To see this, consider an example. Suppose we start with the most minimal conception of action: to act is simply *to bring something about*. On this interpretation, the paradigmatic case of action has the following form: I desire some end X, I see that I could get X by doing Y, so I do Y. Action aims merely at effecting a change in the world, so that the world conforms to my desires.

Mill endorsed this conception of action. As he put it,

> All action is for the sake of some end, and rules of action, it seems natural to suppose, must take their whole color and character from the end to which they are subservient. (Mill 2002, 2)

So, to act is simply to try to bring about some desired end; the rules of action, the standards of success for action, pertain solely to how well the action brings about this end.

Let's translate Mill's point into the terminology of constitutivism. What is constitutive of action is simply what is constitutive of bringing about ends. Is there anything that is constitutive of bringing about ends? Well, a condition on bringing about an end

is taking the necessary means to that end. Thus, if you aim to bring about an end, you must aim to take the means to that end. So:

(1) An agent's ϕ-ing is an action iff in ϕ-ing the agent aims to bring about some end.
(2) An agent aims to bring about an end iff the agent aims to take some of the necessary and available means to this end.
(3) Therefore, an agent's ϕ-ing is an action iff in ϕ-ing the agent aims to take some of the necessary and available means to her end.

From (3), it follows that in each token of action, the agent aims at taking the necessary and available means; it also follows that aiming at taking the necessary and available means is part of what makes an event an instance of action. By the definition of Constitutive Aim, this is just to say that the taking the necessary and available means is a constitutive aim of action.

But now consider the instrumental principle: if an agent aims to E, and M is a necessary and available means for E-ing, then the agent has prima facie reason to M. Notice that we can derive this principle from the conjunction of Success and claim (3): claim (3) entails that agents constitutively aim at taking the necessary and available means to their ends; Success claims that if an agent aims at X, she has prima facie reason to X. Together, these claims entail the instrumental principle.

In short: we can derive the instrumental principle from a very minimal conception of action.[77] We might describe this as *Humean constitutivism*. As we saw above, Humean views maintain that all reasons are derived from facts about our contingent aims: I have reason to ϕ only if I have some motive that is suitably connected to ϕ-ing.[78] They face the question of why—what grounds this commitment? One possible answer is the argument above: taking the means to one's ends is a constitutive aim of action.

Thus, a Humean version of constitutivism would claim that there is only one universal normative principle, directing us to take the means to our ends. The content of these ends is given by our contingent, subjective motivational states. Humean constitutivists would therefore deny that any substantive universal content follows *merely* from facts about agency. They maintain, instead, that in order to derive substantive content, we need to appeal to contingent and variable facts about the motives of particular agents. So there will be no universal reasons derived from facts about the nature of agency; all reasons will be subjective. The variation in reasons will be as wide as the variation in motivational states between agents.[79]

[77] James Dreier (1997), Christine Korsgaard (1997), and Michael Smith (2012) defend similar claims.

[78] There are two complications. First, some Humeans reject the idea that there are rational requirements of any form, including the instrumental principle. See Millgram (1995) and Hampton (1998). These Humeans will want to deny Success. Second, some claim that the instrumental principle expresses a rational requirement rather than a source of reasons: see, for example, Broome (1999). The idea, here, is that the instrumental principle governs combinations of attitudes, telling us either to give up our end or to take the means to it. I will address this issue in the next chapter.

[79] If some particular motivational states are ubiquitous, then the reasons derived from these motives will also be ubiquitous. For example, sympathy might be widespread enough to generate nearly universal reasons;

At the other end of the spectrum, we have *Kantian constitutivism*. Korsgaard's version of constitutivism posits both a version of the instrumental principle (Kant's hypothetical imperative) and the Categorical Imperative as constitutive of agency. This is a maximally ambitious version of constitutivism: it aspires to show that the entirety of our current moral code can be extracted from facts about what is constitutive of agency. As Korsgaard puts it, "Enlightenment morality" can be derived from facts about agency (1996b, 123).

So a maximally ambitious version of constitutivism claims that the entirety of "Enlightenment morality" can be extracted from what is constitutive of agency, whereas a minimally ambitious version claims that only the instrumental principle can be so extracted. This leaves plenty of room for more moderate views, which would fall between the two extremes. Velleman's theory is a good example. He denies that we can extract our moral code from facts about agency (Velleman 2009, Chapter 5). But Velleman does think that we can extract some substantive normative content, including various norms that "favor morality without requiring or guaranteeing it," from facts about agency (Velleman 2009, 149). In short, we can get more than the instrumental principle but less than Enlightenment morality.[80]

But there is a complication: it's not just the *content* of the constitutivist views that distinguishes them. There are also differences in the characterization of the constitutive feature that purportedly yields this content. Humean versions of constitutivism focus on the presence of a constitutive aim, as I explained above. But a more properly *Kantian* version of constitutivism would view mere *aims* as inadequate for generating normative content; aims, along with associated motives, are (so the Kantian story goes) external to the will, so acting upon them would result in heteronomy.[81] Thus, when we turn to Korsgaard's Kantian version of constitutivism, we find reliance on a somewhat different constitutive feature: not a constitutive aim but a constitutive *principle*. I will explain this point in Chapter 4, when we consider Korsgaard's Kantian theory. For now, it simply bears noting that the Kantian version of constitutivism is based on the idea that action constitutively involves commitment to certain principles. The normative principles that govern the will are the principles that are constitutive of action itself. So, the Humean constitutivist shows that actions has constitutive aims, and appeals to Success in order to derive substantive normative content. The Kantian constitutivist, by contrast, argues that action requires commitment to constitutive principles, and derives substantive normative content from the agent's commitment to these principles.

alternatively, if long-term self-interest were universal, we could appeal to Hobbesian considerations to justify certain reasons. My point is simply that the nature of agency itself won't underwrite these conclusions.

[80] Michael Smith (2012) briefly (and tentatively) argues for a rather different version of constitutivism, which centers on the claim that rationality requires that people intrinsically desire that there is as much happiness as possible.

[81] Kant writes, "if the will seeks the law that is to determine it anywhere else than in the fitness of its maxims for its own giving of universal law . . . heteronomy always results" (*Groundwork* 4:441). I will address this issue in Chapter 4.

In sum, constitutivism is a label for a broad range of views united by their attempt to ground normativity in facts about what is constitutive of agency. There is conceptual space for many different versions of constitutivism: we can be Humean constitutivists, Velle-manian constitutivists, Kantian constitutivists, and much else besides. In the following chapters, I will be arguing for a version of constitutivism that—like Velleman's view—falls between the Humean and Kantian extremes. This will be a Nietzschean constitutivism.

4. Conclusion

In this chapter, I have suggested that the familiar ethical theories—non-reductive realism, Aristotelianism, Humeanism, Kantianism—have difficulty overcoming the epistemological, metaphysical, and practical challenges. Each of the familiar theories has trouble with at least one of these challenges. A constitutivist ethical theory, by contrast, would overcome the challenges with ease. For constitutivism extracts norma-tive content from the structure of agency, by showing that all actions share a common aim. In deriving normative content from this aim, it has the potential to avoid the practical and metaphysical problems; and in giving us a way to assess our current moral beliefs, it can avoid the epistemological problem.

But can we develop a successful version of constitutivism? The theory is ambitious, attempting to extract normative content from the bare idea of agency. In light of this, it has been subject to a number of seemingly powerful objections. The next chapter examines and assesses these objections.