Chapter 4: Revisionism
***Four Views on Free Will, 2^{nd} Edition* (in press)**
Manuel Vargas

1 Changing Our Minds

We haven't always seen the world as we do now. There was a time when people thought water was one of the four basic, indivisible substances of the universe. Virtually no one thinks that now. For large parts of the 19th and early 20th century, many people educated in Europe and the US were convinced that race was a biological category. Today, perhaps the standard view is that race is mostly a social category only loosely linked, if at all, to biology. For centuries and perhaps millennia, many people thought whales were fish. Today, we know that whales are mammals and not fish. Some will go on to add that, anyway, the category of fish is, itself, something of a mess.

Science is full of examples of relatively radical transformations in our understanding of things, but this isn't just a scientific phenomenon. For example, philosophers have sometimes found plausible broadly revisionary accounts of psychological attitudes, personal identity, and gender. Nor is this phenomenon restricted to academics. Some people used to insist – and some still do – that marriage is a sacramental relationship between a man and a woman (and God, perhaps). However, ordinary usage, the historical record, and in many places, even the laws, now operate with a very different understanding of marriage. In short, we – scientists, theorists, and laypersons alike – have changed our minds about the nature of a lot of things.

This chapter defends the view that, with respect to free will, we are as people once were with water, race, and marriage. We have a familiar, widely recognized way of thinking about free will that admits of more and less elaborate theories. Yet, even our best theories tend to encounter oddly recalcitrant intuitions and conflicting convictions. In some moods, free will seems to require powers that seem extraordinary, given what we know about the rest of the world. At other times, it can seem obvious people have free will. When we deliberate about what to do, it seems that we must assume we are free; when we have been wronged by anyone of normal ability and maturity, we tend to think they acted freely and culpably, unless they advance some excuse. Yet at other times, it seems unclear why we even need the concept at all. Perhaps it is a relic of a fading religious framework, or an illusion generated by a deceitful psychology that encourages us to see the world as filled with magic.

There is no shortage of proposals for how to explain these puzzles, tensions, and challenges. If you've been reading this book, you have already seen three of the very best proposals for how to understand free will. Each chapter represents a rich tradition of efforts at unravelling the puzzle of free will. Yet, for all the centuries (millennia, really) invested in resolving the problem, none of the standard views has secured anything like a consensus. A conjecture I explore in this chapter is that an important difficulty we have in theorizing about free will is rooted in the shared assumption that the target of our theorizing must be free will as we have tended to conceive of it. On the standard approach, to produce a theory of free will, we must thus first identify its essence. Yet, this effort seems entangled in the persistence of the debate: most disputants are sure that the free will at stake is the free will we imagine ourselves to have, but there is robust disagreement about what exactly that free will comes to.

One way out of this quagmire is to resist hanging the success of our theory on identifying a conceptually nonnegotiable essence that shows up in all the diverse ways we think and talk about free will. That effort to identify essences is fated to founder on the fact that our everyday intuitions don't admit of a unified, coherent story. Our ordinary concept (or, if you prefer, our cluster of sometimes tacit commitments about the nature of free will) is too disordered and tendentious to admit of an elegant resolution in a theory wedded to its details. Free will matters for our practical and social lives, but a satisfying theory of it cannot vindicate the tangle of convictions that have, over the centuries, sprung up around our thought and talk about free will. A satisfactory theory of free will is going to require a new understanding of free will, one that captures many central elements of what is at stake in our concern for it, while abandoning other elements that, on reflection, are less important than they have seemed in our ordinary thinking about free will and its stakes. I call this picture "revisionism about free will."

The version I defend holds that free will exists, it is compatible with the possibility of determinism, and its distinctive features are a function of its mediating various practical and social interests. It falls short of the metaphysical aspirations some of us have for free will, but it nevertheless explains why what we have is free will worth the name.

2 Kinds of Theories

Consider the distinction between how we do, in fact, think about a target concept – whether "fish" or "foul" or "free will" – and how we ought to think about it. In aiming to capture how we do think about the nature of a thing, we are giving a *diagnostic* account. In aiming to defend an all-things-considered judgment about how we should be thinking of this thing, we are giving a *prescriptive* account.

To see why this difference matters, imagine Athena is one of the first chemists, in an era where people tend to hold that water is one of four basic and indivisible substances. Athena's proposal is that water is H_2O, something composed of hydrogen and oxygen. In that time and place, her theory would be revisionary because it conflicts with then-ordinary views about water. Suppose, though, that she wanted to anticipate how much resistance her theory would meet. To do that, she would need to know the widespread (if confused) thoughts about water had by her community. For that task, she would need a different kind of a theory, a theory about people's ordinary understanding of water. This latter thing is something we can call a "diagnostic theory" of water. A diagnostic theory won't tell us what, all things considered, we ought to think about water. Still, diagnostic theories can be useful for understanding why people say and do the things they do. Having an accurate diagnostic account can help us predict why and on what basis people might resist the new chemical theory of water.

Having an accurate diagnosis of ordinary thought is sometimes vital. Suppose that someone named Barrows rejects biological theories of race but is regarded by people in her time as being a member of a biological race that is held to be dangerous, unpredictable, and morally inferior. It might well be a matter of life and death that she has an accurate model of how people around her think about race, even if she regards it as a bad theory of race. Barrows would have the burden of needing what W.E.B. Du Bois (2017) called "double consciousness" – a good model of other people's values, beliefs, and habits of interpretation, even if she repudiates some or all that picture. This is what a diagnostic theory provides. Yet, neither Barrows nor Athena needs to think that an accurate diagnostic theory constrains the best theory of race, water, or, for that matter, free will.

Some theorists reject the need, appeal, or even the possibility of revisionist theorizing, at least in some contexts. They are often motivated by the presumption that conflicts with ordinary beliefs is evidence the theory has gone wrong, that it is changing the topic, or that it has incurred a significant theoretical drawback in departing from ordinary convictions. Nonrevisionists favor *conventional* theories, or theories that do not conflict with ordinary convictions about some topic. A conventional account of something will have more bells and whistles than layperson views, but those details are intended as coherent developments of the basic architecture of everyday commitments about that thing. Unlike revisionary theories, conventional theories tend to have very little daylight between their diagnostic and prescriptive accounts. Most accounts of free will are conventional theories.

What of free will skeptics or eliminativists? They hold that we lack the things required by our diagnostic theory of free will, so we lack free will. (I will treat “eliminativism” and “skepticism” as interchangeable. Some use “skepticism” to refer to views that are only dubious about free will, reserving “eliminativism” for views that reject free will, holding that we ought not employ it in our truth-seeking talk about the world.) Skeptical views are typically conventional theories in the relevant sense: they hold that a satisfactory or correct theory of free will (even a theory that concludes that we do not have it) must cohere with the commonsense features of our diagnostic accounts.

Revisionary theories in any domain are often born of the insight that we are not limited to the verdicts of our diagnostic theories. If philosophical or scientific study reveals that we lack some feature that figures in our thinking about free will (or water, or marriage, or race, or…), this does not necessarily doom the possibility that we have that thing. If there is a successful positive prescriptive proposal, one that explains what free will is in a way that is illuminating and sufficiently continuous with enough of our relevant thought and talk about free will, then we can insist that we have free will.

Until relatively recently, the possibility of revisionism has been mostly overlooked in scholarly discussions of free will. Yet, a revisionary theory of free will is not just an abstract possibility. It is an especially powerful approach for explaining both what is appealing about standard theories and why, despite their venerable pedigrees, none of those more familiar theoretical approaches has succeeded in producing the kind of convergence of scholarly opinion we expect from successful theories.

If I am right that no standard account of free will can coherently capture all the conflicting aspects of everyday thinking about free will, then there is a trivial sense in which every account of free will is tacitly revisionary. However, a theory of free will is explicitly revisionary if it proposes an account of free will that takes itself to conflict with commonsense views about free will; such a theory proposes abandonment of some element in ordinary, widespread convictions about free will. Going forward, I employ this sense of revisionism about free will. A credible revisionary theory requires an explanation for why revision is called for. It needs an account of why the proposed revision(s) to our thought, talk, and/or practices constitute an improvement or advance over alternatives, including those that reject the existence of free will.

There are many ways to be a free will revisionist, and concrete proposals need to be evaluated in all the usual ways. That is, revisionary theorizing is still subject to evaluation in terms of explanatory power, parsimony, theoretical fit with our best understanding of the world and of free will's apparent importance within it. At the same time, the free will revisionist must justify the theory's departures from common sense. The revisionist's not-so-secret hope is to contribute to a transformed understanding of the subject, so that what is initially reviewed as a revisionary picture eventually comes to be regarded as the new default understanding of that thing (at least among theorists, if not always among laypersons). In the end, how radical the proposal is partly depends on the context. If the proposed revision is only a modest departure from the convictions of a few, the revision is minor. If it is a significant departure from the convictions of many, it is a major revision. For any account, it is ultimately an empirical question how revisionary it is.

My account is revisionary about the concept of free will because I think important strands of ordinary thought are as libertarians have said. A variety of considerations – including the power of philosophical arguments in the vein of the Consequence Argument, experimental findings about people's conceptual commitments, and the long arc of our cultural history – suggest that many of us are at least sometimes committed to a form of agency that requires indeterminism. To put my cards on the table, I think many people – maybe most – have an earnestly held picture of free will according to which metaphysically robust notions of sourcehood and leeway are required of free will, at least sometimes and in some contexts, and they think we have these powers. By "metaphysically robust," I mean that these notions require things like nondeterministic causation, perhaps grounded in some nonreductive, emergent feature of agents according to which the action isn't entirely explained by features of the world existing prior to and external to the agent's deliberation or choice-making. Crudely: wide swaths of folk beliefs are libertarian.

There are, of course, many ways philosophers have tried to vindicate those convictions. A common strategy is to speculate that perhaps ordinary physical things, when arranged in the right ways, produce novel causal powers that do not reduce to their components. Free will leverages the happy accident of human beings having such powers. A different strategy attempts to explain how humans might harness existing physical indeterminism in a way that doesn't require positing a novel form of causation. Kane's account in Chapter 1 is often thought of as the most promising instance of these strategies. It relies on the idea of neural quantum amplification, holding that the brain can be in a state that makes it sensitive to quantum effects. A still further strategy holds that perhaps there is an order of things outside the physical, causally ordered world that can act upon the physical, causal order, and that is where free will is located. Immaterial souls and Kantian noumenal selves are accounts in this vein. Nothing in what follows depends on choosing between these regimentations, for that is what I think they are. Libertarian philosophical theories are attempts to make sense of everyday convictions that are oftentimes inchoate or unelaborated, but sufficiently contentful to exert pressure on what we regard as a satisfying account of free will.

On standard approaches to free will, we try to construct an account of free will by looking for its essence. We start with some supposedly neutral demarcation of the subject matter (e.g. "the ability to do otherwise," "the control condition on moral responsibility," "...in the basic desert sense"). We then elaborate on the general idea. On this picture, the conceptual work is front-loaded. We test the contours of our ordinary convictions by considering thought experiments, building arguments that rely on intuitively appealing principles, and checking for the fit of our verdicts with ordinary thought and talk. If a proposal runs afoul of our antecedent sense of the meaning or nature of free will, that's a cost to the theory. Finally, we check for fit with the world. We might find we are free and responsible, or that our belief in our freedom was in error, or that we are trapped in an inescapable but real illusion about our powers, and so on. The whole project, though, depends on that first step of accurately identifying some conceptual essence that is taken to set the constraints on theorizing.

The approach pursued in this chapter doesn't try to identify some essential conceptual content had by all free will thought and talk. Instead, it looks for the cognitive and/or social function of free will. As with any bit of theorizing, it first requires some coarse-grained demarcation of what we are looking for. However, unlike the alternative approach, it puts to the side questions of what is conceptually essential. Instead, it focuses on identifying what free will thought and talk *does* for us (i.e. how it functions in our cognitive and social lives). This can include normative functions like grounding justified blame.

With an account of free will's function(s), we then look to the phenomena of the world. We ask whether there are things in the world that produce, rely upon, or have similar functions as those we associate with free will. If we find none, we join eliminativists in asserting that free will does not exist. However, if most or all the identified functions can be found in the world in a relatively unified way, then we conclude that we have found what we are looking for, even if it lacks some of the features that figure in our everyday thinking. On this picture, locating the doings or functionings is the crucial step, not the identification of the full content of our naive convictions.

Putting things this way can make it look like the deck is stacked for success. However, the methodology of functions over conceptual essences provides no guarantee that we will find what we are looking for. We might find only partial or fragmented bits of functioning that do not constitute a relatively unified thing. Or we might find nothing at all. Our current understanding of the world is one stripped of sorcerers, absolute time, and the divine right of kings. There were no good candidates with comparable functioning that fit with our best understanding of the world. So, we might yet conclude that free will cannot earn a place in our understanding of the world, even if we are prepared to take seriously the possibility that our best understanding of it may be a revisionary one.

When we stop looking for essences and start looking for functions, the result is a happy one. Put one way, this is entirely unsurprising. To borrow a phrase from Kant, much of our modern understanding of the starry heavens above and the moral law within has come about precisely because we moved away from focusing on *whats* in favor of identifying *hows*. So it goes for free will, too. When we go looking for how we would need free will to function and what in the world functions in that way, we discover a real power that makes sense of free will's importance and role in our life. Even better, it fits with a broadly scientific picture of the world. The result is a deepened understanding of free will, one that gets us what we were looking for while simultaneously showing how we were able to talk meaningfully about it, despite stubbornly false beliefs about its nature.

3 Diagnostic Remarks

In Section 2, I claimed that (i) ordinary or "folk" thinking about free will has libertarian elements to it that cannot be vindicated and (ii) that my positive account will not attempt to rescue those elements of ordinary thought. Given that the distinctive mark of a revisionary theory is its putative conflict with ordinary convictions, one might reasonably press the question of why we should think that folk thinking has libertarian strands.

Here, I appeal to the standard arguments advanced by incompatibilists, including those canvassed in Chapter 1–3 of this volume. Elsewhere, I have made my own case for "folk libertarianism," but there are no knockdown arguments to be had by anyone. Still, all my account requires is that a significant number of people have earnestly held commitments about the nature of free will that are most naturally interpreted as incompatibilist, and that in having these commitments they are not misunderstanding their commitments. The extent to which these commitments are widespread among us, where the "us" is, minimally, most members of Western, educated, industrial, rich, and quasi-democratic nations, is an empirical question. Still, libertarians are on to something. They are picking up on a real thread of widespread self-understanding. For some, these commitments may be superficial and readily excised; for others, though, they may be so deeply entrenched that no theory of free will can seem adequate without them.

Why adopt a revisionary theory of free will, one that proposes an account that conflicts with our broadly libertarian commitments? There are three interlocking considerations in favor of revisionism: (i) the implausibility of folk commitments; (ii) their gratuitousness with respect to what is properly central to our thought, talk, and practices bound up with free will; and (iii) the overall advantages of the revisionary alternative. That is, although there are libertarian strands in ordinary thinking about free will, it turns out there is a compelling basis for abandoning them. The powers central to our practical concerns are what we really ought to have in mind when we talk about free will. Importantly, this is a discovery that comes from looking at the actual roles played by free will thought, talk, and practice, and not by armchair reflections on free will's meaning and intuitive metaphysics.

As it was with marriage, race, and many other concepts, we can learn important things about them by looking carefully at what such thought, talk, and involved practices *do* in our lives. For free will, I argue that at the end of that process, what we find is that even though many people believe it is transparently obvious that free will requires powers characteristic of libertarian accounts, those intuitions – powerful as they are – have no more authoritative role in formulating our best theory of free will than our, uh, fishy former convictions about whales. This turns out to be a good thing, as even the best libertarian accounts face serious challenges.

Even if I am mistaken about there being widespread strands of incompatibilist thinking in everyday thought and talk about free will, there is still reason to consider the positive proposal about free will advanced in this chapter. If the positive account proves to be more illuminating or plausible than alternative accounts, there would be reason to accept it even if I am wrong that swaths of ordinary thinking about free will are libertarian. Of course, I do think ordinary thought about free will, and related notions like moral responsibility tend to have incompatibilist strands woven into them. I also think recognizing this fact is important for explaining the persistent difficulty philosophers have had in resolving the debate. It also matters for avoiding objections that have been standardly advanced against accounts that do not think of themselves as revisionary, and that thereby tend to leave undiscussed the special burdens of proposing a revision in our concept and, perhaps, our practices.

My theory of free will is intended to be revisionary. A successful case for a revisionary theory of free will must do at least three things.

First, it must provide a new, positive, or prescriptive account of free will, of how we *ought* to understand what it is. A case for revision – as opposed to the elimination or flat rejection of free will – depends on there being a sufficiently appealing alternative to the difficulties of our pre-revised convictions. That alternative needs to be robust enough to allow us to see how things might look with a reformulated picture of free will, and to consider its account of the roles that free will plays in our lives.

Second, we need to understand why the proposed account of free will counts as a theory of free will, as opposed to a verbal trick or a change in topic. We don't make progress in understanding water or marriage or anything else by simply stipulating that we are going to use the words and their associated sounds to pick out something entirely disconnected from what we had been talking about.

Third, there must be some reason to favor the revisionary proposal. If it does no better than our existing concept, there is little reason to undertake what will inevitably be a slow, effortful reshaping of our thought, talk, and practices.

In what follows, I argue that free will exists, that we often have it, that it plays important roles in our lives, and that it can be further investigated, including with a range of empirical and philosophical tools. What makes the account revisionary is that it conflicts with some important aspects of common, although perhaps not universal, ways of understanding free will. To motivate this picture, I'll focus on what may initially seem like largely tangential considerations about our social practices. If I am right, though, these considerations turn out to be relatively central to seeing our way to a better theory of free will, one that explains why it seems to matter whether we have it, and why there is so much disagreement about what it consists of. I'll then conclude with a discussion of why this account, or one like it, does better than conventional or nonrevisionary theories of free will.

4 Anger

Here are some truisms about free will: free will has something to do with deliberation about what to do; it involves the ability to act; it is the kind of thing that sets us apart from most of the rest of the natural order, including many or all animals; and our having free will is what makes sense of the idea that we can be morally faulted for what we do, at least some of the time. We can get a theory of free will that does all of this, but the best way to get there is to postpone questions about what kind of concept or essential properties can do all that, and instead, to consider some interesting features of our social and emotional lives.

Famously, many people have thought our emotions are in some way fundamentally at odds with rationality. Strongly felt emotions tend to make people fixate on the source of the emotion, often distorting their judgments about that thing. Anger, especially, motivates people to seek retaliation or the imposition of the loss of some value on the offending party, often at great cost to the one seeking it. Sometimes that cost is even greater than the initial harm or loss that motivated the anger. In those cases, even if the good is recovered there is still a net loss. Thus, strongly felt anger can seem like a fundamentally irrational way of responding to the world. If the emotions are so detrimental to rational calculations, and to our securing the goodies that enable us to live and to live well, how did anger and other strong emotions come to be part of our psychological repertoire?

A collection of psychologists, anthropologists, and behavioral economists have converged upon an illuminating answer: anger and other strongly felt emotions are often useful over time and in groups *because* of their disposition for motivating costly action. Imagine a society of people who are frequently self-interested but never prone to anger or other strongly held emotions. If I know you are a purely rational deliberator, not disposed to anger, then I know I can steal from you, break contracts, or otherwise cheat you, just so long as I am confident that it is more costly for you to recover the goods than to let me go. After all, it is going to cost you if you must shut down your shop to chase me or if you might have to hire other people fast enough to chase me down. Unless the chance of recovery is sufficiently high and the value of what I cheated you of was substantial, spending resources to go after me would be to risk throwing good money after bad. Given that you and I are both rational, I know that you won't go after me if I make stopping me just risky or costly enough to you. So, we both know I can and will exploit you with some regularity. I can come by every day and do the same thing, repeatedly, and every day you won't expend the effort to stop me because it is never rational to do so in any serious way.

Notice, too, that in a society where everyone shares this psychology, trust is hard to come by. Agreements aren’t worth the paper they are written on and what coordination and cooperation there will be is going to be a matter of whether there is enough overlapping immediate self-interest to make it work before the risk/benefit ratio of defecting from cooperation shifts.

Let’s change the story, though. Suppose that as you are restocking an item I (again) stole from you the day before, you bang your head and lose consciousness. Upon waking you find that you now experience strong feelings about a wide range of things. One of those feelings is anger. Reflecting on the way I’ve systematically taken advantage of you day after day, you feel the flush of something you will later identify as rage. You feel a strong need to do something about my mistreatment of you. You promise yourself that things will change.

The next time I come into the shop, you are ready for me. You’ve hired a security guard. He costs you more than the value of what I steal over time. Still, you now think that maybe you need to stand up for your interests. After all, your dignity and self-respect have value, too. The next time I come around for some low-key theft I see the guard and barely make it out of there. That turns out to be a prelude to something even more unexpected. Normally, once I make it out of your store, I can be confident that it is irrational for you to chase me down when there has been minimal loss. Normally, I can break into a quick run for a few seconds and then you give up. This time, though, the guard keeps following. Even more surprisingly, you leave the store and start shouting to anyone that will listen that you will pay an absurdly large bounty for anyone to get me. My light and formerly very rational theft has become way more trouble than I bargained for.

The injection of anger-capable psychologies changes the risk/reward calculus in a dramatic way. A society with members generally disposed to anger at offenses isn’t just a society inclined to a strong emotion, or a set of strong emotions. It is also one where agreements and elaborate forms of longer-term social cooperation become possible in a way that wasn’t available for nonangry, self-interested agents. It isn’t that it is impossible for the balance of self-interested reasons to favor breaking social norms. However, the possibility of anger-motivated confrontation and the demand that its targets suffer some loss of goods or interests (whether rights, privileges, or well-being) give social norms a much more powerful basis than local arrangements of nonangry rational self-interest.

Importantly, anger doesn’t only do its work in the case of being angry at someone who directly harmed me. Perceived harms directed at my family, clan, or community can stir me to anger on their behalf, even if I don’t witness the harm, and even if I am unclear about the identity of its perpetrator. In addition to this complexification of our outward-oriented psychology, there is an inward-oriented change, too. If I fail to live up to norms I accept, I can become angry at myself, and I can think it appropriate that I suffer a setback of some of my interests. What all these forms of anger share is that they are agent-directed responses to wrongdoing. Getting mad at the sky for ruining your day with rain does little to alter how the sky treats you, but getting mad and withholding offerings to the sky god might. In short, we have pressures for having anger that is a response to perceived wrongdoing of other agents, and that anger motivates us to either seek confrontation or to impose costs in response to that action.

This is obviously a toy example. We almost certainly didn't get emotions by bumping our heads. Moreover, the story as it tends to be told by some of its proponents tends to be evolutionary. On that approach, the acquisition or development of retributive emotions (i.e. "backward-looking" or "for the record" reactions according to which wrongdoing merits the imposition of a cost or loss of interest on the wrongdoer) was a product of natural selection, an outgrowth of more basic reactions that turned out to be beneficial for the survival of those with this trait. It is a picture supported by its continuity with animal psychologies, and various formal models suggest advantages for communities that have "costly" norm enforcement of the sort implied by retributive anger.

What my account requires is that we have these attitudes, and that they tend to produce these effects. Nothing depends on their having evolutionary origins. The issue is how our social psychologies function at scale. There can be many individual cases of anger or other reactive emotions that don't produce the relevant result so long as we get the right results overall. This is an important feature of functional explanations more generally. If your lungs fail to function in a specific instance – as when the wind is knocked out of you, or when you get a lung infection – this does not speak against the truth of one's lungs having the function of helping you breathe. Similarly, that retributive anger might have the function of bolstering norm enforcement, and relatedly, enabling stable forms of cooperation and coordination, is compatible with there being many cases where it fails to do so. Similarly, that people are not aware of the function, nor have it in mind when they are so functioning, is no argument against the truth of a functionalist account. People can be entirely unaware that they have lungs, or what the lungs are doing, all while enjoying the benefits of functioning lungs.

Anger doesn't preclude the possibility of all wrongdoing, but it does make it more costly. It enables people to have greater confidence in widely known expectations, rules, and norms because everyone knows that others are disposed to enforce those rules, even at high cost to themselves. The stability of those rules encourages social trust, and it enables more complex forms of coordination and cooperation, which in turn facilitates longer-term planning, which in turn secures certain kinds of personal goods (e.g. becoming good at a team sport) and collective goods that require scale and time (including aqueducts, sewage, and organized food safety measures). In short, we have a first idea: agent-directed anger contributes to the stability of coordination and cooperation.

5 The Control Rule

A second and independent idea can help us appreciate how some subtle conceptual and social innovations can improve life in a social world that is shaped by anger and what P.F. Strawson (1962) called "the reactive attitudes." Let's return to our fictional world, wherein the advent of angry agents changes the incentives to engage in exploitation and cheating. Living in a world of relatively bare-bones angry retaliators has a new cost, though: everyone is chronically at risk of triggering other people's wrath. Anger raises the cost of violating norms and agreements, but it also makes us vulnerable to anger-motivated retaliation.

What produces this new cost is the coarse-grained nature of the bare disposition to strike back in anger against some source of harm. It produces what is sometimes called a system of *strict liability*, where penalties are insensitive to the transgressor's reasons or intent. In a strict liability or strike-back-based system of anger, if Amalia's chinchillas destroy your epazote patch, you'll be mad at her even if she took every reasonable measure to keep them away from your epazote. Similarly, if Michael violates our local dance norms by stepping on Taylor's toe during the village dance-off, that he did so unintentionally makes no difference to strict liability punishers.

Plausibly, there have been communities where anger practices have been, and perhaps sometimes continue to be, indifferent to the reasons, motives, or degree of control of offenders. But there is a threefold cost to organizing social practices around our coarse-grained emotional dispositions to automatically seek retaliation for any violation of agreements, norms, or expectations.

First, our lives become unpredictable, subject to seemingly capricious setbacks to our interests. It simply doesn't matter whether I did what I did unknowingly or by accident, or, for example, whether there was some even worse fate lurking if I didn't so act. It is therefore exceptionally difficult for us to anticipate and avoid the risk of angry retaliation for norm violations.

Second, in a world of bare-bones anger-as-commitment-device, there is no in-principle basis for determining when an offender has paid enough. Bare anger's systemic magic is that it is relatively indifferent to other value trade-offs, and apart from exhaustion, there isn't obviously a principled stopping point at which angry retaliation becomes too costly or inappropriate to pursue. Recall the wrath of Achilles for Hector's killing Patroclus: Achilles re-enters the Trojan War, knowingly forfeiting his immortality to seek not just the death of Hector but also his elaborate dishonoring.

Third, there is the cost of interpretive disagreement. Even when we have a relatively robust awareness of what the norms, agreements, and expectations are, we might still disagree whether they apply in this case. The upshot – reciprocal violence – is familiar from history and literature: if a Capulet retaliates against a Montague for violating some norm, but the Montague disagrees that he violated that norm, then the Montague will seek to retaliate for that unjust retaliation by the Capulet, and so on.

To sum up, in a world structured by bare anger at perceived wrongs, we get to enjoy the otherwise surprising possibility of stable cooperation, enhanced coordination, and the possibility of longer-term planning. Yet these goods are vulnerable to the volatility of the emotions involved. Liability is unpredictable, anger can be unconstrained, and interpretive disagreements threaten our access to the collective goods of cooperation and social coordination. We can close the gap between a practice like that and our more familiar understanding of a system of responsibility, culpability, and norm enforcement by introducing something we can call a "control rule."

The issue is this: we have an independent, self-standing interest in controlling our exposure to having our interests set back. In this context, that means we're interested in managing our liability for unknowingly violating a norm, even unknowingly. One way to do this is to agree on a new rule: *don't retaliate against wrongdoing if the wrongdoing was not in the wrongdoer's control*. This new rule gives us a way to manage our exposure to the risk of liability. If Ximena does something by accident, or if she was unaware that she was violating the rule, or there was some special consideration that made the rule violation the right thing to do all things considered, then Ximena doesn't have to be worried about whether she is going to be subject to retaliation because of her inadvertent violation of the norm.

Incorporating a control rule in our retributive, norm-enforcement practices unlocks many of the key features of our contemporary understanding of moral responsibility, including excuses, restricted targets, and various refinements to our system of holding one another responsible. First, consider how adoption of a control rule brings with it the possibility of an excuse. Recall that the control norm precludes retaliation where things weren't in the wrongdoer's control. Where you lack control, you cannot rightly be subject to the angry reactions that would otherwise follow. This is an excuse. An excuse blocks culpability for some piece of wrongdoing, and it is an important feature of our contemporary responsibility practices.

(Perhaps excuses can be arrived at without a control rule, but having a control rule ensures we have some notion of excuse. Again, the point here isn't to tell the actual history of how our practices developed, but to identify important ways our responsibility system functions.)

On this picture, the notion of excuse is the complementary shadow cast by control – or, perhaps, control is excuse's complement. Either way, it is in this reciprocal relationship between control and excuse that we begin to zero in on free will, including what it is and why it matters. I'll come back to that in a bit. First, though, I want to draw your attention to a general framework for thinking about control, something we might think of as an *ecological* approach to control. On this picture, there are at least three crucial elements of an individual agent's control: the ability to recognize considerations that bear on the agent's aims, the ability to guide one's behavior in accord with the considerations that bear on those aims, and the presence of situational circumstances conducive to so acting.

If I cannot see that there is a crocodile coming to eat my child, my lack of knowledge precludes my having control over whether my daughter is going to be eaten. If I see there is a crocodile, but I am gripped by seizures as I try to save her, or I am unknowingly paralyzed by something I ate for lunch, then I lack the ability to control whether my child is going to be eaten. If I see there is a crocodile, and I have control over my body, but I am separated from my child by some impediment like quicksand, or an insurmountable barrier, then the environment precludes my having control. Each of these constitutes a different possible source of excuses. We might think of these as cognitive impairments, volitional impairments, and situational impairments, respectively. If one faces sufficient impairment along any of these axes, then we may recognize an excuse.

On this picture, it is a mistake to think we can say all that we want about agency, or the ability to act, without thinking about how an agent's abilities interact with an environment. Two agents with identical physical configurations might face very different challenges if their environments differ. My having poor vision is mostly immaterial given how our physical, material world is now arranged. My myopia would have been a very serious issue on the savannah of our ancient ancestors. For creatures like us, control is always ecological in the sense that it is a function of the powers of agents in circumstances.

Some of those circumstances are sociohistorical. Plausibly, communities differ about what counts as an impairment sufficient for excuse. Perhaps some communities do without excuses along one of these axes, or only recognize excuses in especially extraordinary situations. Depending on what considerations and varieties of control we emphasize, we might cultivate both *peoples* and *environments* so that they have widely variant capacities for meeting those standards. Twenty-first-century Silicon Valley is likely to demand, and to train, configurations of cognitive, volitional, and environmental control that are distinct from those of Carthage in the third century BCE. Even so, both communities will have an interest in fostering people who are sensitive to considerations of harm and to the importance of social norms, and who are responsive to the general conditions of cooperation in everyday social life.

So, a control rule gives us a notion of excuse. A control rule also enables a second feature characteristic of our contemporary system of moral responsibility: target selectivity, or some notion of being an apt or proper target. To see the importance of this idea, recall that in a simple system of anger-motivated retaliation, you might be able to get decent levels of norm enforcement from retaliating against a range of possible targets. I might deter your future norm violations by retaliating against your family or your clan. In a more narrowly control-responsive system of responsibility, though, there is new pressure to restrict the anger to offending agents. After all, it defeats the purpose of a control-responsive practice if you are unexpectedly on the hook for something someone else did. The foregoing thoughts tend to bring into sharper relief a question that might lurk from the outset of any responsibility practice: to whom do responsibility practices apply?

Today, we tend to think young children, those afflicted by various infirmities, and nonhuman animals aren’t suitable targets for the full range of ordinary anger-propelled attitudes. Yet, different groups have had different views. For example, restricting our attention to the European tradition, it may help to know that *mens rea*, or the “guilty mind,” requirement that we think of as central to contemporary criminal law came into existence in the 12th century. And, as late as the 18th century, animals were put on trial in European criminal proceedings. Although the history of the criminal law is not the same thing as the history of responsibility practices and norm enforcement more generally, it is not unrelated either. That history serves as a reminder of different ways communities have settled the question of proper targets, or conversely, who is exempt from responsibility practices.

Here’s the form of a general answer to the boundaries question, though: we restrict our norm violation reactions to those suitably able to recognize and respond to norms, or to those who recognize the reasons reflected or expressed in those norms. This might seem only to postpone the hard question. What counts as a suitable ability to recognize and respond to norms and their underlying reasons? There are different answers one can give here. For example, one might appeal to the bare susceptibility to any improvement in control, sensitivity to moral considerations, or dispositions for prosocial cooperation. On that picture, if you can or could come to be positively affected by angry blame practices, you count as a proper subject of angry norm enforcement. A worry, though, is that this forfeits the virtues of prediction and control that were the point of moving away from a coarse-grained strict liability approach. Young children or animals might not apprehend the norms at all but might still be moved in the right ways by angry blame practices.

A different account, and one that I prefer, locates the answer in the connection between a person’s interests in being held responsible and the interest we have in holding that person responsible. Ordinarily, we have a reason to be seen as competent at navigating the normative demands of our communities, including demands on how we conduct ourselves, reason about things, and so on. We want to enjoy the statuses and privileges of those who are regarded as fully mature, sufficiently competent members of a community that depends on relatively complex norms of cooperation and coordination. So, any failure to adhere to those norms creates a problem for managing one’s reputation.

We solve the reputation-management problem by giving some strong signal of our competence with those demands even when we fail to meet them. In short, we acknowledge the normative failure and accept responsibility. The details of this can vary depending on the norm and the perceived harm. However, in accepting fault, we recognize the suitability of negative judgments and our potential liability for repair and apology, precisely because we share in the judgment that our control was good enough for those norms to apply to us. The otherwise curious phenomenon of our willingness to take the blame for our errors is thus explained by our commitment to the norms, the communities for whom these norms matter, and our standing interest in being regarded as competent at respecting those norms and their reasons.

On this picture, then, suitable control is, roughly, control sufficient to sustain some default level of coordination and cooperation in societies that have responsibility norms at all. In different times and places, we might draw the line for suitable control in distinct places. Still, something like a Goldilocks principle seems right. It can't be so demanding that no one satisfies it, because then we lose the benefits of having a responsibility system in the first place. Yet, it can't be so flexible that everyone satisfies it, or else we lose the predictability of being vulnerable to blame that was the appeal of a control-sensitive practice.

6 Adjusting the Rules

The advent of a control rule gives us a picture of excuses and generates some pressure to provide an account of the scope of responsibility practices. It also points to a third way further innovations can take shape in our practices, over and above the simple picture of bare anger. Once a community recognizes the possibility that it can institute rules about when it is apt or permissible to express and act on anger, it can continue to adjust those rules. Excuses might be regarded as a matter of degree, and we might develop nuanced pictures about the varieties of different mental states and their implications for culpable wrongdoing.

We might, for example, decide that running a risk of violating a norm (recklessness) is not as bad as intentionally violating a norm. We might also expand the ways one can avoid liability to anger by, for example, introducing ideas that certain kinds of choices can be unreasonably difficult to resist (coercion and/or duress). Along a different dimension, we might narrow the way excuses work, seeking to block strategic claims of ignorance about vital matters by introducing the thought that there are things any competent member of our community ought to be sensitive to, regardless of whether it springs to mind (so, a negligence rule).

This picture suggests two other advances that are worth mentioning. Recall the general problem of volatility that besets a system of bare anger. Beyond the fact of unpredictability, which is relatively directly addressed by a control rule, there are two other recurring sources of volatility in an anger-based practice of norm enforcement: unconstrained retaliation and interpretive disagreement. Both phenomena create the risk of unremitting reciprocal violence, where Capulets relentlessly retaliate against Montagues for a harm that the former think is profound but the latter reject as mistaken or minor.

We might, however, exploit a feature of the control rule to establish a kind of constraint on expressions of anger and its disposition to fuel confrontation and retaliation. The operative idea is to limit the degree or kinds of expressions of anger (including sanctions or other ways of imposing losses to one's interests) by the degree of both wrongfulness and control exhibited in the offending act. This is the idea of proportionality. Our response to wrongdoing is apt only up to the point to which it reflects something about the moral quality of the wrongdoing. A great wrong that was not culpably done (because of excuse or exemption) does not license great retaliation. A great wrong without excuse or exemption (or some explanation of why it was the right thing to do) can license a proportionately great deal of angry norm enforcement.

A norm of proportionality, even in conjunction with a control rule, would leave a fair amount of volatility in this newer, more sophisticated social practice of angry norm enforcement. The reason is immediately recognizable: even if we agree about the necessity of control and the principle of proportionality, we might disagree about the nature of the wrong and what counts as proportional. The form of a solution is equally familiar: some mechanism of independent adjudication. Some forms of adjudication can be institutional, with relatively defined roles for resolving conflicts (such as duels and judicial proceedings). However, interpretive disagreements can also be solved in nonformal but deeply social ways. We appeal to a friend or a mentor to get perspective on our interpretation of events. Sometimes, we even appeal to a wider, entirely impersonal group of strangers to assess matters – plaintively asking in the frank language of this era: AITA?

Communities aren't just passive bystanders, springing to action only when asked. When there is convergence about the meaning of actions and norms of proportionality, it seems likely that individuals and groups will be proactive, for example, endorsing one or another interpretation of some act, or insisting that some instance of blame, retaliation, or punishment has gone too far. Conversely, there might be calls that a victim of the initial wrongdoing has not stood up to some offender. Woe to bad interpreters if they are unresponsive to the collective judgment of their communities.

Thus far, the account has focused on the social dimensions of the practice, and the general pressures that operate on it. Yet practices do not operate in social space, untethered from our psychologies. Those practices both express our underlying attitudes, and they also give shape to them. Even if it is not available to us to entirely abandon retributive anger, as some have thought, this does not mean that we cannot come to regiment the nature of its expression and our attitudes to it. There is some reason to think that constraints on *expressing* anger might lead to widespread pressure to disfavor *having* anger when the wrongdoing is excused or justified.

Here's one argument for thinking a community might be inclined to expand the norm constraining expressions of anger at those excused or justified to the repudiation of having angry attitudes in those cases at all: given that (i) being angry increases the likelihood of expressing anger, and given that (ii) expressions of anger toward those offenders with an excuse or justification for their wrongdoing will be regarded as itself a violation of the new norm, then (iii) there will be some reason to view one's own experiencing of anger in such cases as inapt, and thus (iv) individuals will have some reason to acquire methods for suppressing or controlling the experience and intensity of inapt anger. Over time, these pressures would presumably come to affect the way members of that community might educate and train the moral dispositions of the young. To the extent to which such psychology-shaping efforts succeed, those adjustments would make it easier to live in accord with the operative social norms concerning the propriety of anger.

To sum up: having a system of angry responses to norm violations enables forms of coordination, cooperation, and long-term planning that would otherwise be unstable or impossible to sustain over time. However, a simple system of bare, angry responses to any norm violation is subject to considerable volatility. A solution to this problem is to introduce a cognitively demanding innovation, a control rule that calibrates our risk of liability and the degree of that liability in a way responsive to our control over the resultant norm violation. There are different ways that calibration can go, but the centrality of it emerges when we think about its multidimensional significance. A notion of control gives communities with angry norm-enforcing practices a principled and effective way of mediating between (i) collective interests in the goods of angry norm enforcement, (ii) individual interests in demonstrating that one is competent at local norms and a good cooperator, and (iii) ongoing interests in mitigating one's exposure to liability. Finally, we've seen why a practice with this feature would give us reason to shape the sensibilities of those subject to such a practice, treating as a matter of important socialization the identification of various norms about what is important, how norm violations are to be responded to, when anger is apt and inapt, and ways of adjudicating disagreements when they arise.

Again, the aim in this section has not been the One, True, accounting of how we came to have our specific configuration of morally and psychologically nuanced practices of social regulation. Those details are interesting but mostly beside the point for present purposes. Instead, the ambition has been to call attention to an interlocking set of psychological and social functioning that we already enjoy. In focusing on them, we can see how those things create a demand for a kind of control that involves recognizing and responding to normative considerations that matter for complex forms of coordination and cooperation.

That is, the overarching logic or functional structure of the responsibility system depends upon – and is systematically entangled with – the nature of our emotions, the persistence of our interests in the goods of sociality, and the pressures to cultivate the powers distinctive of human agency. These considerations interlock with our interests in prediction and in avoiding setbacks of our interests. Jointly this collection of considerations create pressure for a practically useful form of control, the shape of which is given by those practices, the entwined psychology, and our interests in the goods of complex forms of cooperation and coordination.

7 What About Free Will?

One might worry that something has gone wrong, for free will has not yet made its appearance. But free will is just what I've been discussing. Free will is the ability or capacity for norm-sensitive control suitable for life in a community of creatures engaged in complex forms of coordination and cooperation. This is a higher-order capacity, supported or constituted by a collection of finer-grained recognitional and volitional capacities that enable us to meet the relevant normative demands. That higher-level capacity for recognizing and responding to norms and their reasons – which can be very diverse at the more granular level of the interaction of brains, bodies, and environments – has an identifiable general functional structure involving reasons-responsiveness in particular contexts. So, adding a bit more nuance, we can say that *free will is the situational ability to suitably recognize and appropriately respond to relevant normative considerations*.

Each major functional component of free will (the recognition element, the volitional element, and the situational element) might be realized in diverse ways. For example, your ability to recognize that a loved one is experiencing stress might rely on sensitivity to tone of voice, the posture of a body, the ability to imagine or compare past episodes to the present one, and so on. The ways in which agents might avoid acting on bad reasons might involve distraction in this case, focus on some other good in another case, and in yet another, massively culturally scaffolded efforts at habituation.

With respect to the partitioning of situations (i.e. when is something a similar situation for the assessment of control?), things are similarly flexible. The relevant distinctions presumably start as a matter of local practice sensitive to common contexts and general values. This yields a picture where we with these norms partition relevantly similar situations this way, even if you people do it that way. As groups interact, and the advantages for shared norms relevant to cooperation and coordination do their work, pressures to converge emerge. If a society or group of societies come to recognize that initially distinct-seeming situations share relevantly similar features, they can come to conclude that two putatively distinct situations are a species of the same general situation. They might draw similar conclusions if two situations realize or frustrate values in the same way or are best navigated by people with similar agent-level features.

There are several ways this picture can be developed. One might construe free will in terms of sensitivity to normative considerations, full stop. Or one might restrict it to those considerations that matter for coordination and cooperation. Yet another approach anchors free will in the idea of sensitivity to specifically moral considerations. This last version of free will yields a somewhat narrower band of powers, those organized around the specific role that morality plays in our life. So, for clarity, we might distinguish between wider and narrower senses of free will.

I am inclined to privilege a wider sense, although it isn't unreasonable to privilege or even exclusively employ a narrower one focused on morality, especially considering the central role that concerns for blameworthiness and responsibility have for free will. However, the rational, norm-sensitive contextual capacities implicated in those practices turn out to matter across a wide range of domains, including in attributions of credit and blame in contexts as diverse as art, sports, and epistemic endeavors. What unifies these things is the relevance of "oughts," or normative considerations more generally. Once we have an interest in the oughts required for social coordination and cooperation, there is some pressure to think the relevant kind of control involves any oughts that bear on deciding what to do. So, I favor a very wide notion of free will.

Even if it turned out there was a sense in which the origins of our concept and its role was exclusively rooted in specifically moral practices, this is compatible with our current interest being more capacious than that. Undoubtedly, some exercises of our distinctive, norm-sensitive capacities can have moral significance in some of those otherwise nonmoral contexts. Still, it isn't obvious that every choice in those contexts must be marked by moral significance. If so, that's a reason for favoring a wider notion of free will over a narrower one, even if a diversity of approaches is compatible with the general approach.

What about determinism, indeterminism, reductionism, and all the other "metaphysicalisms" that threaten freedom? On this approach, free will's metaphysics – like many other relatively high-level phenomena – has a functional characterization defined by a cluster of social interests and practices, but a realization given in the physical states that enable that functioning. Consequently, free will is relatively insulated from the particulars of how those functions are realized. The place to look for free will (and felonies, world-class forwards, and fair social arrangements) is at the level of human interests and practices. Only with an account of their functioning in hand does it make sense to see what arrangement of physical things produce those functions. Neuroscience, chemistry, and physics can help us fill in the details of how that functioning works. Yet, whatever the general facts are about the causal microstructure of our high-level functioning, those are details that explain how and not whether we have free will.

This picture is compatible with the idea that free will has a special causal role, despite its being realized in an open-ended set of lower-level physical processes. Suppose everything is linked in a complex but ultimately deterministic chain, where one can trace causes all the way back to some initial event. For creatures like us, with interests like ours, it is nevertheless often important to be able to identify special places in that chain. Speaking metaphorically, we want to know where the hinge points are, whether different kinds of inputs produce varied (even if potentially deterministic) outputs. We have reason to care about those places where we can intervene, places where the causal powers are distinctive, or where certain kinds of information or arrangements of the world can shape which way the hinges are pointing.

To get a sense of why this matters, even in deterministic systems, consider a video game. The typical video game is set up to function as an entirely deterministic system. Even the random elements are themselves often only random-seeming products of a deterministic algorithmic process. Still, as players in the game, we can be deeply invested in whether we are triggering this or that deterministic process. Will that monster be distracted if you put food in its path? Can it be distracted by in-game noise? If so, the monster has hinges. If the monster automatically goes after you no matter what else is happening, then it doesn't have a hinge.

One might protest that if determinism is true, it is also true that our attention and interests are, in part, functions of prior states of the world, whether we realize it or not. Fair enough. This doesn't undermine the significance of hinges, though. When we deliberate about what to do, it is vitally important to us (at least, given that we are hoping to be successful at playing the game) to be clear about which combination of button or keypresses produce which result, determinism or not. The point is that even when we know something is deterministic (as in most video games), we still have a live interest in figuring out how that system works. That knowledge allows us to intervene on the world in ways that are more responsive to our (perhaps deterministically produced) desires, values, and ambitions.

On this picture, free will – that situational capacity to recognize and respond to normative considerations – is itself a crucial human hinge point. It is the place in the causal chain where distinctive kinds of information make a difference in people's behavior and the (at least epistemic) possibilities for how people relate to that information. If creatures have the right kinds of hinges, we can relate to them in the complex, norm-sensitive ways that we expect of ordinary, mature agents. If they lack these things, we don't. (Unless, of course, they culpably made themselves that way.) Free will matters, in part, because it is a central hinge for distinctively human ways of relating.

Why then have so many of us come to think that free will requires something more than some form of competence with normative considerations? There may be no single or simple answer. It may be a byproduct of an easy-to-make cognitive mistake, perhaps involving a too-hasty interpretation of our phenomenology, or a tempting but unmotivated shift in the kinds of explanatory demands we put on instances of action. Or perhaps it is a product of later cultural accretions whose effects linger long after the impetus for adding them disappeared. This is a matter for further inquiry. In the absence of some compelling reason for thinking we need a more demanding notion of responsibility, though, the reality of free will is secured by the fact of our ability to navigate often complex normative demands.

8 Abilities

Here one might protest that this is still too quick, for even on the present account, free will is a kind of capacity or ability. Yet, one way determinism challenges free will is by showing that there is something suspect about the idea of abilities. Perhaps what determinism shows is that the only thing that can happen is what does, in fact, happen. If that's your thought, then because my proposal relies on some construal of capacity, ability, or the idea of "can," you might worry this account is an elaborate cheat, distracting us from the core problem for free will.

In reply, I agree that we can specify modal notions (i.e. a notion that involves the idea of possibility or necessity) such as "can," "ability," or "capacity" so that if determinism is true, we can't do anything other than what we do. That's not the issue, though. Every day we make use of a wide range of more and less demanding conceptions of the idea of "can." For example, if we want to know whether someone can speak a language, catch a ball, or tell whether someone else is sad, we don't need to know whether the underlying features of the agent are deterministic. We just need to know something about the relevant dispositions they have, and whether and how those things function in the range of relevantly similar circumstances at stake in the question.

We make these distinctions in nuanced and scalar ways, reflecting a range of intersecting interests. An alcohol-impaired dancer may be in less control of what he's doing than a sober one, but an impaired concert pianist might have a greater ability to play a passage of Schumann's "Concerto in A minor" than a beginning pianist. Whether Satya speaks Japanese might be something we settle in different ways depending on the context. Competence sufficient to navigate everyday social situations is different than competence at live translation in an academic context. Pinning down the precise specification of any modal notion is difficult. Even so, in everyday life we manage to navigate these distinctions, often with great nuance.

Given the foregoing, the issue (whether determinism means we lack the normative capacity at stake in the proposal of this chapter) depends on whether our having the relevant capacities requires the falsity of determinism. There is decisive reason to think it does not. To see why, consider the idea of abilities that depend on the interaction of the physical features of the agent with the roles defined by a practice. Judges can have the ability to settle a legal case, but the bailiff doesn't; that ability can be disrupted if the judge is sick or not at work. A given athlete might be comparatively capable of scoring in a basketball game because of her skill and knowledge of the rules, but she might be a bad bet in a kabaddi match on account of both her ignorance about the rules and her lack of relevant skills. For these everyday notions of ability, metaphysical truths about determinism are simply orthogonal to whether people have these abilities – at least, in the ordinary, everyday sense.

Even philosophers who believe determinism undermines free will tend to acknowledge that determinism does not undermine the point of deliberation (i.e. settling what to think or do, given the fact of "for all we know" possibilities), that we make choices, that we form intentions about what to do, that we sometimes act for reasons, and that there is a meaningful sense in which people's capacities can vary. The substantive issue that separates their views from the one on offer is whether these everyday notions are sufficient for making sense of our free will thought and talk. The argument of this chapter is that a situational capacity for recognizing and responding to normative considerations is indeed sufficient.

Whether determinism is true or not, everyday distinctions about capacities carve important parts of our lives at their socially significant joints, in part because the powers that matter are those in some sense defined by or dependent on features *internal* to a relevant practice. Questions about who deserves to be fined for speeding are a matter of rules internal to legal practices. Even if there is a standpoint available to us external to this practice, or even one external to all our human practices, it doesn't mean that claims about practice-dependent abilities aren't true, or that they lack authority. We might have reasons to wonder whether we want a given practice, and correspondingly, whether we care about the abilities specified by the practice. However, those thoughts are *external* to the practice, and would need to be anchored in our interests and reasons just as much as anything else. The fact that such a standpoint may be available to us does not vitiate the fact that we can and do have practice-specified abilities and interests.

This is not to deny that practice-dependent notions of "can" interact, or that practice-dependent abilities can rely on yet other notions for some purposes. For example, we might think that, with respect to a given practice, only goalkeepers can use their hands. We might also think that *this* goalkeeper lacks the ability in *this* game because she just came back from getting surgery on both her hands. Again, these kinds of claims are obviously true and important for our everyday life, and again, the status of determinism has no bearing on them.

What then of the modal notion(s) involved in free will? The argument of this chapter has been that there is an important, practice-dependent role to be played by a notion of ability that is responsive to normative demands, that involves recognizing and responding to considerations, that is paradigmatically involved in deliberations about what to do, and that makes sense of the distinctiveness of human beings and the basis of culpability practices. Elsewhere, I've offered a detailed account of what that comes to (Vargas 2013). You don't have to call it free will, full stop, if you don't want to. (My coercive powers over you are minimal.) Still, this notion captures the functions we noncontroversially want out of free will, and it does so in a way that illuminates how and why free will matters, and how and why we can come to have false beliefs and persistent disagreements about the nature of free will. That's good enough for revisionism about free will, though. Any onus is on those who want more.

Again, this is compatible with an interest in highly specified, very demanding versions of "can" questions. We may want to know whether, holding fixed the entire history of the universe and any governing laws, an agent could do otherwise. That we can specify a power that is that metaphysically demanding doesn't show that this is what free will must be; that such a power would be desirable doesn't show that it is necessary. The considerations advanced in this chapter show that the notion of capacity we need for free will is a much looser or flexible one, one that functions in the right way for shared, cooperative life. It is a less demanding picture than our intuitions and default conceptual content about free will tend to suggest.

9 Normative Authority

On the present proposal, free will exists, it is compatible with the possibility of determinism, and it has a functional structure that depends on its role in mediating various practical and social interests. Its metaphysics might seem less glorious than many of us prefer, given its dependence on merely physical and social phenomena, but revisionism about free will is not intended to capture the full contents of our collective imaginations about what free will is like.

Still, one might think that something has gone wrong in this account in a different way. Even if we grant that it identifies a cluster of agential powers that rightly figure in our everyday practices, even if it grounds a web of recognizable statuses and bases for differences in interpersonal assessment and reaction, and even if we allow that it employs a defensible conception of abilities, we might still think that the account is the wrong kind of account of free will. That is, it provides us with various pragmatic considerations for having free will thought and talk, but it doesn't really touch the morally serious core of our interest in it. There are lots of things that are instrumentally useful to us (e.g. pants, passports, and paychecks) that we do not think of as generating the robust normative authority (or the "oughtiness," as the philosopher John Doris has said), that somehow seems implicated in our having free will. So, one might think, something has gone wrong.

Although some people think that morality is just about considerations of cooperation and coordination among creatures with psychologies like ours, I agree that we want free will to do more than support social regulation. We want it to support moral and perhaps other robust forms of normative authority. Having free will is not like being genteel, having a terrific fashion sense, or being adept at platformer video games. The significance of those things is mostly a matter of personal preference. Instead, free will is a power that is intertwined in our collective moral lives and in our own self-regard. It is the kind of thing that seems to ground fault-finding and demands that one account for one's behavior, in a way that is more than a contingent form of social organization. Is there a way to explain that aspect of free will's significance for us?

Yes. Here's the idea: pressures for social regulation give us reason to identify the kind of control at stake in free will, but once we have that idea it turns out to have a further, distinctively moral significance: it is a morally valuable form of agency, one we have reason to cultivate. Recall that the situationally relevant power to recognize and respond to normative considerations is central to a practice of coordination and cooperation. An important subset of normative considerations are moral considerations, considerations that are grounded in what moral reasons there are. (I won't try to say anything about the content of those reasons – that's the subject matter of normative ethics.) The important thought is this: our having free will, and thus our being able to recognize and respond to moral considerations, is a way of having a distinctive and morally valuable form of agency.

We don't get that agency for free. A good deal of childrearing is about the shaping of values and concerns. This tends to be done by identifying for children the ways in which conduct makes one liable for moral praise and blame. We feign blame to teach children when and where they will be genuinely blameworthy when their baseline abilities to recognize and respond to moral considerations is sophisticated enough to meet default expectations required for mature members of the community. The development and eventual achievement of a free will that is sensitive to moral considerations, and that can anticipate and avoid liability for norm violations, is the centerpiece of how we shape our agency. On this "agency cultivation model" of responsibility, free will is at the center. Our participating in responsibility practices is our way of shaping our agency in morally valuable ways. That is, moral values anchor this form of instrumentalism about responsibility. So, even if we get to free will by social regulation pressures, once we have it, it turns out to be of tremendous moral significance in ways that go beyond its utility for social cooperation.

So, cooperative pressures give us a notion of control that in turn specifies the abilities that constitute free will. However, once we have such a notion, it turns out to matter morally, because it provides a basis for responsibility practices. When those control-sensitive responsibility practices are justified (i.e. they do sufficiently well at producing beings with sensitivity to moral considerations), they fix the truth conditions for culpability, blameworthiness, and desert.

Alas, there is no guarantee that fully justified practices are the ones that are operative in a given community. Indeed, we can predict that communities can do better and worse at tracking what moral considerations there are. Thus, that we have free will doesn't guarantee that we do a good job of tracking blameworthiness in everyday life. Still, free will is at the center of that web of practices that make us into morally sensitive agents we have reason to be. So, free will matters morally, and not just as a tool for social regulation.

10 Realism

The present account of free will, with its somewhat socially dependent nature, might raise the worry that it lacks the objectivity or realism that some might expect of free will.

As I will use the terms, a realist about *X* thinks there are facts about *X* that hold independent of us, our reactions, or our attitudes about that thing. So, a moral realist holds that there are facts about morality, and that they are independent of us in some important way. For the moral realist, morality has a nature independent of our thoughts and feelings. In contrast, the moral antirealist thinks that what facts there are about morality (if there are any) depend on how we think about morality, or on our psychological dispositions for responding to things. So, if one holds that a divine being sets up the moral facts about the world before there are any humans or other sentient beings, that's a realist picture of morality. In contrast, a paradigmatic antirealist picture holds that what facts there are about morality are at bottom a matter of how our psychological dispositions and feelings tend to converge into ways of seeing and feeling. For the antirealist, morality is not a matter of our tracking something independent of our psychologies.

My account appears to be antirealist about free will and/or related notions like responsibility, deservingness, and so on. It holds that free will is the kind of thing whose features are partly dependent on human interests and social practices. Undoubtedly, the most natural reading of this account is antirealist. It holds that free will is a bundle of capacities that realize a particular social role (or set of social roles). On this account, free will's nature and significance is emmeshed in a web of statuses and interests that are "post-social," or products of human social lives and our relationship to it. (Post-social does not mean post-institutional; we can have sociality without institutionality.) The account does not invoke some human attitude- or practice-independent notion of freedom, control, desert, and the like. This can seem troubling, though, if one thinks that a theory of free will should hold that free will is independent of our social attitudes and practices, or if one thinks free will requires realist normative phenomena.

Although the account is most naturally read as antirealist, I do not think that it requires antirealism about free will, responsibility, morality – or even for normative and evaluative properties more generally. First, there is the possibility of overdetermination: perhaps there are separate practice-dependent and practice-independent (realist) bases for free will. Second, one could think that a realist notion of free will partly grounds or explains the practice that produces the notion of free will I have identified. Recall that my account requires that our social practices direct our concern toward a bundle of capacities that can perform specific social roles, and it holds that those considerations are sufficient for identifying a set of properties adequate for the various roles imposed by the practice. One might hold that a fuller explanation reveals that the capacities around which our practices have come to gravitate were already there. That we already had abilities with that general shape partly explains why our practices are what they are.

Such a view would not be without challenges. Were one's realist picture of free will to invoke significantly different powers than the account I have offered, one might think that we ought to adjust our practice-dependent notion in the direction of the realist notion. Still, this suggests that this account is compatible with some versions of realism. A third possibility is this: we could conclude that thought and talk about free will are ambiguous, picking out two distinct things, each independently sufficient to justify free will thought and talk. So, accepting revisionary antirealism about one sense of free will need not conflict with a potentially more ambitious further realist sense of free will.

I find this fourth possibility appealing: even if inclined to moral realism, one needn't be a realist about all aspects of morality. One can be a *patchy* realist, accepting realism about some values and not others, or about some but not all normative notions (e.g. aretaic qualities, valuable states of affairs, deontic notions, fairness, etc.). Some normative and evaluative notions may be foundational, with others derivative or dependent on those foundational notions.

As a conceptual matter, responsibility and attendant notions do not seem foundational to morality. We can conceive of a system of morality, as such, without responsibility, culpability, and desert. (Pereboom, at least, agrees.) In contrast, other normative or evaluative notions (including, perhaps well-being, better and worse, and fairness) seem more foundational to the possibility of morality at all. In sum, despite its antirealist appearance, the present proposal is compatible with a range of possible normative realisms. Even if we accept antirealism about responsibility, it's compatible with realism about morality.

Last, notice the account I have offered has an answer to a version of what is sometimes called the "Euthyphro dilemma" for normative realisms. The dilemma is this: either the response-independent thing is somehow suitably connected to us, our natures, and our interests, or it is hard to see why that entity generates reasons for us that aren't capricious or arbitrary. (The dilemma is named for the take-home lesson of a challenge Socrates raised for Euthyphro's claim about piety being what is dear to the gods: there must be something that explains or grounds what is dear to them, on pain of piety being grounded in something arbitrary or without reason.) My account grounds free will in our nature and interests, answering the dilemma without requiring some wider form of metaphysical or normative antirealism. This seems a virtue.

11 Topic Continuity

I favor a form of revisionism about free will that is primarily concerned with how we represent or understand free will. Yet, this sort of change – conceptual change – can also entail changes in the practices that involve that thing. Revisions to our concept of whales, water, marriage, and so on were entwined with a wide range of more or less central changes to practices that appealed to these concepts. On this chapter's proposal, free will is a natural, genuinely robust phenomenon. Yet, it has important limitations. It is a function of how individuals are built, both by nature and enculturation, and it is further constrained by circumstances or opportunities. So, on this picture, free will is a kind of achievement that can be lost, undermined, or only intermittently available, depending on what happens to people and their ecologies of action. I take it that this picture is importantly at odds with some ways of thinking about free will, especially those that treat it as a power that makes its possessors radically independent of the rest of the causal, physical universe. Such a critic might wonder if my revisionary account is still talking about free will.

As noted at the outset, a persistent challenge for any revisionary proposal is to show that it retains topic continuity (or the sameness of topic). If a revision is too radical, it runs the risk of changing what philosophers of language call the referent or target of our thought and talk. When we revised our concept of water, we were still talking about that liquid stuff in rivers and streams. The evidence for topic continuity is that the revisionary proposal captures a wide swath of ordinary thought and talk, it explains how that talk could be true, and it also informatively explains a web of phenomena surrounding that thought and talk. Moreover, it seems to be *true* in a distinctive way, fitting together with other things we know about the world. Similar things can be said of other cases of revision.

The case for topic continuity about free will has a structure parallel to the case for water: in both pre- and post-concept revision, we continue to refer to a distinctive ability or power that enables us to navigate complex normative situations, the possession of which makes our conduct subject to evaluative regard. While there might be some contexts where employment of a revisionary account of free will would constitute a topic change, the proposed account delivers on our central concerns: (i) it captures a wide swath of ordinary thought and talk (i.e. it makes good on the truisms), (ii) it explains how these things could be true, (iii) it informatively explains a web of phenomena surrounding that thought and talk, and (iv) it seems to be true, given everything we know about the world.

No revisionary theory is immune to every objection regarding topic continuity. Consider, again, the old timey theory of water. Suppose members of the Philosophy Club are discussing Empedocles's theory of water because they are interested in understanding the mechanics of an ontology of four basic substances. Suppose someone interrupted the discussion to insist that because water is both hydrogen and oxygen, Empedocles's ontology requires at least five basic substances. The rest of the conversationalists wouldn't be wrong to think that that introduction of the chemical theory of water was a topic change. If one's interest in water is dependent (or, um, downstream) from its being metaphysically basic, then even the (true) chemical theory of water would be a topic change.

Again, the parallel holds. Our topic is what free will really is, or what is getting picked out when we say things like, "You did it of your own free will" or, "It was her free choice to ignore the warning." If, however, someone stipulates a different interest in the term, or a significantly different conceptual role for it, then my proposal may fail to preserve topic continuity.

This account is silent on at least one thing some people have wanted from a theory of free will. If one wants a theory of free will to explain why God is not a jerk for knowingly creating a universe in which infants are stricken with painful incurable diseases, this account comes up short. Or, if one wants a theory of free will to explain why one's wrongdoing in this world might justify eternal damnation, then the present account cannot do that for you. That my account cannot explain why people deserve to burn for an eternity does not seem to me a defect, regardless of whether one is religious.

Let's distinguish between a general theory of free will, whose principal burdens are recognizable across a wide range of usage, and a conception of free will that is intended to serve a particular set of theological commitments. Revisionism about free will is intended as an instance of the former, as an account of what could realize the kind of thing that figures in our ordinary thought and talk about free will. Religious concerns are widespread, and they can impose distinctive burdens on what one wants out of a theory of free will. Yet one can have an interest in free will without being religious. Moreover, given that the religious and nonreligious person can together meaningfully discuss the merits of different proposals about free will, everyone needs a general or nonsectarian theory of free will.

There is a parallel here with marriage. A general theory of marriage must explain a wide range of formal human mating arrangements across times and cultures. A particular religious community might want an "in-house" theory of marriage to do further work, specifying, say, sacramental elements that appeal to a relationship to the divine. Those requirements might be important and proper to their practice of marriage. However, that narrower, sectarian conception of marriage doesn't help us identify the wider range of marriage practices around the world. Thus, a specifically religious conception of marriage needn't be in conceptual conflict or competition with a more general proposal for how to understand marriage as such.

Undoubtedly, many people have reasons to be interested in articulating a religious conception of free will in the same way in which they can have reason to articulate a religious conception of marriage. Nothing here denies the urgency of that project. Yet, from the standpoint of accounting for a nonsectarian notion of free will we need not, and ought not, try to capture the distinctively religious functions that some have asked free will to support.

12 Reconsidering Alternative Positive Views

With a revisionary account in hand, it may be useful to reconsider some of the alternatives. This section focuses on libertarianism and compatibilism; the next on hard incompatibilism.

Many of the thought experiments and arguments standardly offered by incompatibilists have force precisely because they capture real, robust, and relatively widespread features of common sense. Those elements are part of our collective understanding about free will, and they partly explain why there is any philosophical puzzle about free will at all. Yet, there are also widespread aspects of thinking about agency, deliberation, and freedom that lend themselves to compatibilist theorizing. Debates about free will get a lot of mileage out of alternately playing up or disparaging each set of convictions. Effective philosophical presentations tend to encourage their audience to privilege one set of intuitions over others in a process that amounts to a pruning of alternative convictions. In time, those alternative pictures begin to appear increasingly mysterious, disingenuous, or confused. (Perhaps that has happened to you, as you have read this volume?)

If that is right, and if many of us start with libertarian intuitions, at least some of the time, why *not* libertarianism? I see no promising way to vindicate libertarianism in the face of its standard criticisms, even in ingenious and scientifically oriented accounts like Robert Kane's. For all its intuitive virtues, the powers required by libertarianism are metaphysically problematic and normatively gratuitous.

First, it isn't enough to say that a given libertarian picture is *compatible* with the findings of contemporary science. Science doesn't rule out the possibility that ectoplasmic miasmas influence decisions, either. We don't take it seriously as a hypothesis, though, because there is no independent reason to think that such phenomena occur when we decide what to do. Nor do we need it to explain any known feature of decision-making. Absent such considerations, an ectoplasmic theory of decisions is entirely ad hoc. The same is true, it seems to me, of the role of indeterminism in standard libertarian accounts of free will.

Second, and relatedly, libertarianism's scientific credentials are highly speculative. When I put quantum amplification proposals to working neuroscientists, I'm told that contemporary biology doesn't make much use of quantum amplification effects, and anyway, we haven't yet seen these sorts of effects at the temperature that neurons operate. I'm also consistently told that nothing in mainstream neuroscience independently suggests that we have Kane-like amplification of quantum indeterminacy. Parallel concerns arise for versions of libertarianism that invoke distinctive forms of causation, or whose adoption requires that standard pictures of causation are deeply in error.

One might reply that this is philosophy, and not yet science. Future discoveries might yet yield things undreamt of in our philosophies. But the stakes are too high to be indifferent about the epistemic credentials of our practices. We blame, punish, and even kill people on the putative basis that they act with free will. It is manifestly wrongful to punish offenders in the hope that libertarianism is vindicated by future science. This is true whether one appeals to quantum amplification, agent causation, noumenal selves, or anything else so speculative.

In contrast, the kind of power that figures in the present account – the power to situationally recognize and respond to normative considerations – is a readily recognizable feature of our lives. Few seriously doubt that we have it. The free will debate has been persistent in part because we wrongly believe free will must be more than that. The argument of this chapter is that careful attention to the demands of our practical lives and the exigencies of the world show that the power we often have can indeed do what work we want from free will.

Given the foregoing, one might wonder whether we would do better to be conventional compatibilists, or perhaps, semicompatibilists of the sort articulated by John Fischer. What compatibilists get right is that we can identify workable, less metaphysically tendentious notions of freedom that, with some finessing, seem to offer promising bases for organizing our thought and talk about free will, along with related notions like deservingness, culpability, and responsibility. Here, my complaints are less about plausibility than about theoretical significance and burdens.

Conventional compatibilist theories remain in the grips of efforts to respect our theoretical convictions as we find them. Because of the broadly non- or even antirevisionary presumptions of the historical debate, compatibilists find themselves insisting that their proposals aren't in conflict with important strands of common sense, that all they offer are accounts of what we mean ("of all we mean," as P.F. Strawson said) by "free will" and related notions. These are not accounts that intend to offer a revisionary theory, and their proponents make no effort to address the special burdens of revisionary accounts, including responding to concerns about topic continuity. However, to those with broadly incompatibilist commitments, conventional compatibilists are, at best, trying to solve the free will problem with a cheat code. At worst, they are engaged in philosophical gaslighting, denying the force of the intuitions that create the very problem.

Unwillingness to recognize the revisionary nature of conventional compatibilism – perhaps because of earnest beliefs that it is not revisionary – has meant that compatibilists have left unaddressed questions about the basis of the proposed revisions and the grounds for thinking it isn't a topic change. Yet these are precisely the grounds for much of the snark directed at compatibilism, a view that has been described as a "wretched subterfuge," "petty word-jugglery" (Kant), and a "quagmire of evasion" (William James). Unable to recognize revision for what it is, compatibilists have not seen fit to address its burdens. In contrast, an explicitly revisionary approach may be wretched, but it is not a subterfuge. It is an account of what we *should* mean, what perhaps we would have meant, given a better understanding of things.

What about John Fischer's distinguished account of semicompatibilism? It allows that free will *may* be incompatible with determinism, but that responsibility is compatible with determinism. I am more sanguine about the possibility and reality of free will than is Fischer, although we are both unpersuaded by existing forms of libertarianism. One difference between our views is that I think the free will we have is different from the free will we imagine ourselves to have, as I think we have false beliefs about what free will is.

On the matter of the abilities we possess, Fischer and I are in broad agreement. The abilities I have identified constitute free will in the sense that matters; Fischer is less committed to that thought. Elsewhere, I have tried to spell out how I think about moral considerations sensitive agency in much fuller detail, and elements of it have figured in the present account. Unlike Fischer, I tend to think that the kind of "reasons-responsiveness" that matters is at the level of agents, not subagential mechanisms. I have also emphasized a broadly ecological picture of agency, where these powers are indexed to circumstances, and where the degree of responsiveness required might vary by type of consideration and by local concerns. Fischer's commitments on these issues are unclear to me.

Methodologically, our accounts come from different places. I have largely eschewed the conceptualist path of testing intuitions and abstract cases, focusing instead on the functioning of our practices and psychologies as they operate in the context of our ongoing scientific accounts of them. My metaphysics of free will derive from an analysis of our psychologies and the demands of our social practices. For Fischer, the direction of explanation mostly seems to go the other way.

That we come to a similar place is perhaps a point in favor of both accounts. (Or, at least, it is good news for me to be in his distinguished company!) Still, this points to a wider difference in orientation; beyond fit with everyday thought and talk, my account seeks to identify why the powers specified by the account ground the practical and normative authority of our concern for free will and associated practices of accountability. This is, I take it, a less central concern of Fischer’s account. Even so, presumably his account could be supplemented in ways that would more directly address normative concerns.

13 Free Will Eliminativism

Here, I turn to hard incompatibilism. I begin with some general reflections on eliminativist views in philosophy. Then, I consider two elements of Pereboom’s groundbreaking and insightful account: first, the methodology, which purports to fix the terms of the debate by stipulating conceptual content that is supposedly neutral, and second, the content of that stipulation. I argue it is insufficient to rule out the revisionary alternative I have offered.

Eliminativisms in many philosophical domains have faced some common challenges, including the existence of revisionist alternatives and the high costs of error theories. In evaluating the appeal of hard incompatibilism, it may be useful to keep in mind two wider lessons of other philosophical eliminativisms.

First, we now know that in many cases the impossibility of satisfying an armchair concept doesn’t tend to entail that that thing doesn’t exist. (Remember, the world can be different from what we think.) In many cases, we can accept most or even all the premises that animate the eliminativist without accepting the eliminativist conclusion. We can’t justify eliminativism by identifying a merely conceptual defect, or even a failure of the world to vindicate our concepts. So, what the eliminativist owes us is a still further argument, some reason to think there isn’t and can’t be an adequate revisionary alternative. Absent that, eliminativisms are suspiciously close to being a hasty inference that begs the question against revisionism. Of course, if the revisionism doesn’t work, then and only then does eliminativism go live. The upshot is that hard incompatibilism is, at best, a position of last resort.

A second general challenge for eliminativisms concerns the location of the error. Eliminativisms are typically error theories, in that they hold that we are systematically in error about something. For example, the hard incompatibilist thinks that ordinary ascriptions of free will and responsibility are in systematic error. The problem is that the bigger and more sweeping the error claimed by the theorist, the more plausible it is that it is the theorist, rather than the other competent users of the term, who errs.

Still, not all error theories are on a par. Free will revisionism asserts that our convictions have a comparatively modest kind of error, concerning beliefs about the nature of free will. Thus, the revisionist’s call for conceptual housekeeping is less dramatic than the eliminativist's or skeptic's call to radically transform our moral psychologies and social practices. That's because the revisionist's diagnosis of error about *what* some X is is typically less sweeping and problematic than the eliminativist's diagnosis of error about *whether* X exists.

Partly because of these pressures, contemporary philosophy has witnessed a now-familiar trajectory of eliminativist views on topics as diverse as race, propositional attitudes, and the moral virtues. On those topics, as the theoretical options came into better focus, eliminativists tended to retreat from the more ambitious versions of those claims, gradually transforming into revisionists. In short, initially enthusiastic eliminativisms tend to give way to a more sober revisionism.

None of this shows that the free will eliminativist is fated to become a revisionist. Still, the underlying logic across all these cases is that the conceptual troubles that motivate eliminativisms tend to be captured by revisionisms, with less costly consequences for thought and talk. Still, there are no short cuts in weighing out the theoretical alternatives. Some surprising conclusions end up being true. Even if there are general reasons to regard hard incompatibilism with modest skepticism, it deserves to be considered on its own merits. Let's turn to that now.

Recall that, at the outset, I proposed that we understand the nature of free will and debates about free will in terms of what free will talk, thought, and practices do for us. The contrast is an approach that tries to identify some essential content to the idea of free will that any account must satisfy. I take it that Pereboom's approach is an instance of the latter, a broadly conceptualist approach that endeavors to identify some essential feature of free will ("the basic desert sense") that we can use as a basis for conceptual precisification and elucidation. I rejected that approach because it risks building into our accounts contested and erroneous features about the nature of free will. While Pereboom and I agree that there are incompatibilist elements in ordinary thinking about free will and moral responsibility, we disagree about whether an adequate theory of free will needs to be beholden to those elements.

Pereboom's approach for fixing the subject matter of the debate, and for blocking the risk of topic change, centers on the idea of what he calls "the basic desert" sense of responsibility. As he puts it in Chapter 3, "The desert at issue here is basic in the sense that the agent, to be morally responsible in this sense, would deserve the pain or harm, the pleasure or benefit, just by virtue of having performed the action with sensitivity to its moral status, and not, for example, by virtue of consequentialist or contractualist considerations.” So, his target notion of free will is something required to make sense of responsibility with that notion of desert.

By my lights, this is the wrong way to set the terms of the debate. Even if we agree to fix our theoretical target by its relationship to moral responsibility, this approach runs afoul of the injunction to avoid trying to identify essences, to avoid potentially building into our account false beliefs about the nature of free will, responsibility, or desert. (Note: if this chapter is right, our having false beliefs about free will is actual, not merely possible.) In contrast, my functionalist methodology makes no stipulation about the essence or nature of the desert that figures in our practices. On this approach, the proper theoretical target is not a philosopher’s armchair construction of some sense of desert, conjured from one’s intuitions and sense of fit with language. Instead, the relevant target is whatever sense of desert is minimally required to make sense of responsibility practices like these (i.e. the ones we find in the real world), that is *the stuff of actual praising and blaming practices*. That might turn out to be Pereboom’s notion of basic desert, but it might not be.

A quartet of methodological matters merits mention. First, it isn’t clear what the rules of the game are for characterizing basic desert. Since he first introduced the idea, Pereboom’s characterization of it has grown more elaborate over time, adding a variety of additional conditions and exclusions (including the rejection of some bases of justification, the emphasis on pain or harm as opposed to having one’s interests set back, and the idea that the imposition of that deserved harm being noninstrumentally good). Theoretical positions rightly respond to pressures to refinement over time. Still, it isn’t clear why it isn’t enough that the relevant notion of desert at stake in ordinary practices meets only some of these features, as opposed to all.

Second, there is a tradeoff between the degree of specificity in some notion and our ability to assess whether it is indeed the notion that figures in our thought and talk. The more elaborate the account of desert is – and it is now more rather than less elaborate – the less clear it is that this and not some other notion is the best fit with our explanatory ambitions.

Third, it is particularly problematic for Pereboom’s project that we can successfully identify an obtainable notion of desert – nonbasic desert – that makes sense of our existing practices. Consider the nature of penalties in social practices like sports. Typically, a system of penalties is justified by forward-looking considerations, including protecting the health of the players, preserving enjoyment of the game, and creating constraints for innovation within the play of the game. Yet, whether and when something is really, truly a foul is an entirely backward-looking thing, a matter of what happened and what the rules are.

Were we to look to the systematic justification of fouls to assess whether something is a foul, we would fail to apply the rules governing when something is a foul. Of course, we might have all-things-considered reasons to not call some fouls (e.g. if an all-powerful being threatens to end life as we know it if we call a foul) or more pedestrian pragmatic reasons for ignoring the foul call rules (e.g. it is the end of the game, the outcome is settled, and everyone just wants to go home). Strictly speaking, though, whether something *is* a foul (its truth conditions, as philosophers sometimes say) is settled by what the rules say, and not by whether calling the foul in that case contributes to the health of the players, the enjoyment of the fans, or the development of strategic play. In a straightforward and recognizable sense, you only deserve to have a foul called on you if you in fact committed a foul as specified by the rules. Whether we have reason to adopt different rules or all-thing-considered reasons to not call the foul is something above and beyond whether that thing is a foul.

This basic explanatory structure generalizes to a wide range of practices, including the law and parts of morality. Recall the picture I've offered: responsibility practices are norm-enforcement practices that enable coordination and cooperation, whose moral authority, when it obtains, derives from its systemic effects on cultivating people's moral agency over time. The justification of responsibility norms is instrumentalist or forward looking, but the statuses within a practice thus defined are mostly backward looking. That is, the truth conditions about whether someone deserves blame is, within the practice, settled by backward-looking conditions. The content of the rules, and their truth conditions, make no reference to the system-level justification.

The significance of this is easy to miss. Because there are some contexts in which there is no loss of goods if one fails to distinguish between truth conditions and justification, philosophers sometimes run together the issues of the question of truth conditions (when does someone deserve blame?) and the basis for having those truth conditions (what makes those and not some other truth conditions the relevant ones?). A collapse of those questions – sometimes characterized as "tiers" or "levels" (the level of truth conditions; the level of justification) – isn't especially plausible for responsibility. Crucially, a system of responsibility grounded on predictability, information management, the reduction of free-riding, and the cultivation of moral sensibilities *only gets these goods by keeping first-order questions of truth status separate from systemic justification*.

The foregoing thought is about the normative logic, but it is buttressed by our psychologies. Our ability to fluidly engage in responsibility practices depends on our not always needing to ask about whether blaming or punishing will produce the right effects. How would we accurately decide such things on the fly? This is why we do better to internalize relatively coarse-grained norms that are sensitive to apparent control, thereby making characteristic ways of seeing and responding to norm violations a matter of enculturated habit.

A technical aside: Pereboom's characterization of basic desert curiously blends a characterization of desert's truth conditions (that something is deserved because of the nature of the agent and the moral quality of the act) with a set of constraints about the justification of those truth conditions (that they turn on neither consequentialist nor contractualist considerations, which means their effects or significance for hypothetical contracts don't count as relevant to justifying desert). The curiosity is that truth conditions and the basis of their justification are conceptually independent issues; they do not collapse in the context of responsibility. In the context of foundational normative ethics or practical reasoning as such, the truth conditions more plausibly collapse into questions of justification. That's why people have worried about tier collapse for rule utilitarianism, which holds that right action is that which conforms to a rule that would have the best effects overall. However, nonfoundational normative/evaluative practices needn't collapse in that way when the systemic good is secured precisely by having statuses internal to the practice that detach from the systemic justification. That's the lesson of the foul call case. Why then unify these things in a single notion of desert?

Here is a fourth and (for now) final concern. It seems that hard incompatibilism comes to nothing if implementing it leaves our moral practices intact. So, Pereboom's position requires that there be no basis for desert-entailing moral anger. Yet, my proposal identifies such a basis, albeit one that appeals to nonbasic desert. In reply, Pereboom has pointed to Kant's test as proof that this is not enough. If we cannot say that someone deserves blame or punishment even if there is no one else around to administer the punishment, then we do not have the kind of desert at stake in responsibility practices. This constraint is not unreasonable. It identifies a notion of desert without trying to specify its conceptual essence. Tellingly, my nonbasic desert picture readily satisfies it.

Recall that whether someone deserves blame is a function of whether they culpably violated a justified normative rule. (Why have a rule that we blame that violation? Answer: it enables cooperation and builds better beings.) So, suppose someone kills another person. Suppose, too, that everyone on earth besides the offender disappears six seconds after the murder, but before anyone can react to it. On this picture of responsibility, it would still be *true* that the murderer would deserve blame and punishment, even if no one was around to administer it and even if there were no positive effects in doing so. Like foul calls and criminal statuses, one can truly deserve these things even if no one is around to administer them and even if so administering them doesn't produce its customary effect.

The foregoing means that Kant's test of retributive desert can be passed by a nonbasic desert theory. To be sure, this is not a "presocial," from the "standpoint of the universe independent of human life" sense of desert. At best, it is a nonobvious, even surprising alternative basis for retaining our practices. Still, to the extent to which Kant's intuition is the idea that Pereboom is trying to capture with his notion of basic desert, and to the extent to which this is the stakes of his concern in debates about free will, then it seems that by Pereboom's own lights, a nonbasic desert theory of free will gives us what we need to reject hard incompatibilism.

14 Free Will and Back Again

The free will we were looking for wasn't the free will we needed. Like love, marriage, water, and all the rest, sometimes the world doesn't match our expectations. The animating idea of this chapter is that it isn't the precise content of free will that matters. Trying to achieve convergence about the full and specific content of the idea of free will is what has made it so difficult to resolve debates about free will. However, by focusing on the roles that free will thought, talk, and practices play in our practical deliberative lives we can locate a picture of free will that is informative and that leaves room for the possibility that we have erroneous but nonfatal errors in our thinking about free will.

We can agree that an adequate theory of free will should aspire to capture most of our coarse-grained truisms about free will. That's what this account attempts to do, and I have argued that we can capture the relevant truisms: free will has something to do with deliberation about what to do; it involves the ability to act; it is the kind of thing that sets us apart from most of the rest of the natural order, including many or all animals; our having free will is what makes sense of the idea that we can be culpable for what we do, at least some of the time; and free will matters morally. The picture on offer vindicates all these thoughts, without invoking the libertarian powers to which our imaginations direct us.

This revisionary picture comes at a necessary but acceptable cost: it conflicts with elements of common sense, including the idea that we are ultimate sources of what we do, perhaps with metaphysically robust alternative possibilities sometimes available to us. Still, our having free will marks us out as a distinctive part of nature, albeit one not radically separate from it. If things go right – if our biology and social conditions permit it – we often come to develop the distinctive normative powers that serve to mediate between the competing demands of sociality and self-interest. This is a good thing. These powers help us resolve deliberative challenges small and large. Although these powers are defined by a web of practical and normative functions, they are plausibly realized or built up out of thoroughly physical systems.

The foregoing means that we can make discoveries about the underlying physical structures involved in free will, in much the same way we have made discoveries about the composition of water, the diversity of marriage practices, and all the rest. For example, we could learn that other beings have this power in some or another form. Perhaps a few nonhuman animals have the requisite cognitive power to respond to moral considerations in the right ways. Alternately, we might learn that some human pathologies are more disruptive to free will than we thought if, for example, the disorder disrupts the ability to recognize or suitably respond to normative considerations.

It is part of the power of this approach that free will becomes a more tractable empirical matter, something whose contours we might come to understand by engagement with the relevant sciences. We needn't settle those particulars here. It is enough that we have an account that allows us to begin investigating things in those ways. That we can see how they might be answered suggests that we have made real progress.

Further Reading

Despite lingering disagreement about how to regiment labels, a variety of recent accounts have articulated broadly revisionary or potentially revisionary views about free will and moral responsibility. These include McGeer (2015, 2019), Doris (2015a), Nichols (2015), Deery (2021), and McCormick (2022). McCormick (2016) and Vargas (2023) provide general overviews of the literature and the main issues.

This chapter retains the general orientation of the view presented in the first edition (Vargas 2007), but it also contains numerous, sometimes substantial, transformations of my earlier views. The fullest statement of my version of revisionism, including of responsibility-relevant abilities and the agency cultivation model, is in Vargas (2013), with important updates in Vargas (2021). In Vargas (2015), owing to work by McKenna (2009), Pereboom (2014), and especially Doris (2015b), I changed my mind about how to think about desert, thereby abandoning significant elements of my earlier accounts (Vargas 2009, 2013). Wang (2021) has led me to be more explicit about the diversity of social and normative functions of responsibility; work by Brink (2021) and Nelkin (2016) has influenced me throughout.

I've argued for a two-tiered (2007, 2013), desert-invoking (2009, 2013) view of the justification of responsibility practices. Rawls (1955) and Hart (1961) provide the classic articulations of the idea of backward-looking rules justified by forward-looking considerations. Caruso and Pereboom (2022, pp. 9–10) have attributed discussions of desert and two-tiered justifications of responsibility practices to Dennett's monographs (1984, 2003), but I find explicit discussions of those things in those texts difficult to locate. In explaining when "tier collapse" isn't a threat, I've drawn from ideas in Hooker (2000) and Enoch, Fisher, and Spectre (2021).

Forerunners of contemporary free will revisionisms include Smart (1961), Bennett (1980), Walter (2001), Singer (2002), and Arneson (2003). Jackson (1998) self-identifies as a compatibilist but glosses it in revisionary way. Early Dennett (1984, 2003) is sometimes read as a revisionist, although the interpretive issues are complex (see Vargas 2005).

Overviews of recent efforts to study ordinary thinking about free will can be found in Björnsson (2022) and Nadelhoffer and Monroe (2022). For an intriguing, broadly functionalist proposal for explaining the shape of core puzzles, see Björnsson and Persson (2012). Citations to the wider literature on incompatibilism, libertarianism, reasons-responsiveness, and free will skepticism can be found in the prior chapters. Deployments of broadly ecological accounts of normative agency can be found in Morton (2011), Hurley (2013), Vargas (2013, 2017, 2018), and Nelkin and Vargas (forthcoming).

The *locus classicus* of the adaptive picture according to which retributive emotions enable cooperation include Frank (1988) and Fehr and Gächter (2002). For more recent overviews and complexifications, see Nichols (2015, Chapter 6), Cushman (2015), and O'Connor (2022). Many contemporary philosophers are interested in the intersection of naturalistic norm-sensitive agency and dispositions for cooperation, although not always with a focus on evolutionary explanation. Among the many works in that spirit, see McGeer (2012), Zawidzki (2013), Doris (2015a), Bicchieri (2017), Bratman (2022), Kelly (2022), Nichols (2022), and Madigan (forthcoming). My deployment of the idea of causal "hinges" draws from interventionist approaches to free will employed by Roskies (2012) and Deery and Nahmias (2017). For an instructive set of discussion between philosophers and neuroscientists, see Maoz and Sinnott-Armstrong (2022).

Nietzsche's *Genealogy of Morality* is the classic account of responsibility's genealogy. For a recent effort in that spirit, see Pettit (2018); for some challenges, see Plunkett (2016) and Vargas (forthcoming). Sommers (2022) discusses the diversity of responsibility practices. For a model of society-based justified norms, see Copp (1995). For the history of *mens rea* in English criminal law, see Chesney (1939). On animal trials, see Carson (1917) and Cohen (1986). For a contemporary overview of agency in animals, see Monsó and Andrews (2022). Miller (2020) argues that fish are not a well-ordered kind.

Since the first edition of this book, many of the issues that arise for revisionary theories of free will have been explored in an independent but parallel literature, under the guise of "conceptual engineering" and "conceptual ethics" (e.g. Burgess, Cappelen, and Plunkett (2020).