What is the Free Will Debate Even About?

ABSTRACT: Debates about free will are famously persistent. One underappreciated source of the persistence of these debates is a failure to recognize the role of second- or higher-order disagreements, that is, disagreements about what concepts, phenomena, or practices an account of free will is supposed to capture or explain. Another source of the disagreement is a failure to mark a distinction between theories that take themselves to be bound by the general contours of ordinary pre-philosophical convictions about things and theories that are prepared to allow for sometimes notable conflicts with pre-philosophical convictions about those things. A satisfactory construal of the subject matter of free will debates must therefore allow for the possibility of both higher order disagreements and the fact that we might have false beliefs about the nature of free will.

This article argues that a promising way of accommodating these thoughts is to appeal to a broadly functionalist account of free will, one that holds that the proper subject matter of theories of free will is a power necessary to make sense of everyday responsibility practices. In construing the subject matter of debates about free will in this way, we can make sense of both central philosophical debates about free will while also recognizing the possibility of false beliefs about free will's nature and the possibility that a satisfying theory of free will might conflict with ordinary pre-philosophical convictions about the nature of free will.

Forthcoming in *The Harvard Review of Philosophy*

What is the Free Will Debate Even About?

Manuel Vargas, UC San Diego

Theories of free will typically purport to offer an account of what free will is and whether it exists. There are famously long-standing and perhaps intractable disagreements about many of the central issues.

Libertarian theories (theories that hold that we have free will and that this is incompatible with determinism) seem remarkably speculative, postulating powers and pictures of the physical order we have no independent reason to believe in. Moreover, they ask us to praise, blame, and punish on the promise that such powers will be later discovered in some as-yet inexistent science.

Typical compatibilist theories (theories that hold that free will is compatible with determinism) strike many people as an evasion, for they seem to deny the existence and force of intuitions that give us the free will problem in the first place.

Skeptical or eliminativist theories like hard incompatibilism (the view that we lack free will and that it is incompatible with determinism) ask us to abandon wide swaths of our everyday practices with little evidence that it is possible for us to do so, at least collectively in the kinds of contexts we now live. They also ask us to bet that the costs of doing so are not worse than the costs of living with false beliefs about free will and responsibility.

One possible explanation of this situation is that core disagreements about free will reflect a potentially deeper or more foundational disagreement about what a theory of free will is supposed to show. Theorists disagree about what concepts, phenomena, or practices an account of free will is supposed to capture or explain. This essay is about such disagreements and more generally, how disagreements over the proper subject matter of philosophical debates about free will might be resolved.

To address these issues, I begin with some distinctions. I then offer a new proposal for how to understand the subject matter of theories of free will. The proposal is perhaps deceptively simple: the proper subject matter of theories of free will is a power necessary to make sense of everyday responsibility practices.

Distinguishing debates

Suppose it is the eighteenth century, and I am an Empedoclean about water, in that I think water is one of the four basic and indivisible substances of the universe. You are a fan of the new molecular theory of water. You think it is a combination of two different and more basic

things: hydrogen and oxygen. So, we disagree about the nature of water. This is what we might call a first-order disagreement.

Disagreements about philosophical, scientific, and theoretical matters can occur at different registers. While you and I disagree about the nature of water, we might also disagree about several other things that bear on how we understand our first-order disagreement and how we try to resolve it. For example, I could think that a theory of water must clearly preserve my core theoretical beliefs about it. So, for me, a theory of water is under pressure to show that it is one of the fundamental and indivisible building blocks of reality. I might also think that it is a virtue of my theory that it fluidly integrates with the dominant Aristotelian framework of our era. You could reject these commitments. You might think that a theory of water can ignore our favored convictions about the nature of water, so long as it explains something about its qualities, behavior, or relationship to other phenomena. By your lights, it is no theoretical virtue at all that a theory of water coheres with a broader Aristotelian worldview.

These are second-order disagreements. What makes them second-order disagreements is that they are not about water *per se* so much as they are about the basis on which we settle on proposals about water. Examples of second-order disagreements are disagreements about what makes for a good theory of water, what kinds of things a theory of water needs to explain, the basis on which we should evaluate various proposals about water, how we settle disagreements about the meaning of water, and so on.

Many apparent first-order disagreements may turn out to be a cocktail of disagreements at the second order (or beyond). In the water example, we have a first-order disagreement about the nature of water, as well as second-order disagreements about what those theories need to include. The force of these convictions can vary. Sometimes, convictions about second-order questions shape which theory one favors at the first order; at other times, first-order convictions drive the selection of higher-order principles.

The distinction between first- and second-order disagreements is not limited to the phenomena of the natural sciences. There have been and continue to be similar multi-level debates about a wide range of topics, some scientific, but many others social, moral, and political. The resolution of first-order debates about the nature of marriage (whether it is sacramental and restricted to heterosexual unions), debates about race (whether it is biological or social), and more historically, debates about political authority (whether it was a matter of inheritance or popular sovereignty) often require an appeal to higher-order principles.

The immediate recognizability of first-order disagreements sometimes masks more foundational disagreements at the second order. For example, the appeal of the religious view that marriage is exclusively heterosexual and sacramental may depend on a second-order conviction that one's theory of marriage must cohere with the central tenants of a religious

tradition. Similarly, the appeal of a social constructionist theory of race can depend, in part, on both the conviction that there is no adequate biological basis for race, but also that race talk has an important social and historical role that grounds ongoing discourse about race. In sum, the strength of first-order disagreements can depend on one's higher-order convictions about what we are beholden to and what we should value in constructing our theories.

Consider a second distinction, one that cuts across different orders of theorizing. We can distinguish between diagnostic (or descriptive) and prescriptive theorizing. In providing a theory (of race, marriage, morality, etc.) we could give a roughly descriptive account of what people tend to think about that thing, or if you like, the concept or representations people have in their thought, talk, and practices concerning that thing. A different ambition for a theory is to offer an account of what, all things considered, people (whether people in general, or some more restricted group—for example, we theorists, or we for whom those truths especially matter) ought to think about that thing.

Descriptive and prescriptive theories can come apart, and our interest in each can be varied. Long after the advent of the germ theory of disease, people held on to the older miasma theory. Germ theory was plausibly the right prescriptive theory, but in that era the best diagnostic theory of what people thought disease was would have been better captured by a miasma theory. Similarly, even after theorists began to dismiss the divine right of kings as wishful and pernicious thinking, many royalists continued to think there was some special divine authority for rulership beyond popular sovereignty. Depending on our interests, we might care a good deal about what people think, even if it is not what, all things considered, we ought to think. I might have a strong prudential interest in an accurate theory about what people around here think about the basis of political authority if I am considering shouting that both kings and cockroaches ought not exist.

Many philosophers tend to take their projects to be centrally concerned with what we ought to think about some target notion. The methods and evidential basis for such theories can be diverse. The point is only that many philosophers take it that philosophical theorizing aims at telling us something about the truth of the things about which they theorize. In this context, ordinary or "folk" convictions about that thing might be a useful entry point for theorizing. In such cases, I might have second-order convictions that a theory that captures ordinary convictions is to be preferred. Yet those are usually only instrumental considerations, the kinds of things that are taken on board because they are thought to lead us to the right or best theory about the target notion.

The difference between descriptive and prescriptive theories is not always marked in theorizing. Some theories about a given thing collapse the distinction, holding that what we do think about the target idea, concept, or phenomenon is what we ought to think about it. For example, in the thirteenth century many Europeans accepted the view that the Earth

was the center of the universe and other objects rotated around it. In that era, geocentric theorists presumably thought that what the people of that time ought to think about planetary motion was what they already did think about planetary motion. Call this "simple descriptive theorizing."

A related possibility, one more widely held among contemporary philosophers, is that a good prescriptive theory ought to avoid conflicting with what people think about that thing. This is not quite the same thing as simply restating the folk view, for it permits the theorist to add to or otherwise refine the folk theory of some subject matter. Call this "elaborated descriptive theorizing." The ambition of elaborated descriptive theorizing is to respect the commitments of folk theory by not conflicting with its commitments. The contribution of the theorist is in the development of a theory within those broader constraints. To accept this kind desideratum on theorizing is, of course, to accept a higher-order view about what makes for a good theory.

Elaborated descriptive theorizing may seem especially appealing to people who think the principal job of philosophy is, for example, the analysis of concepts, or the analysis of distinctions in ordinary language, or the characterization of the latent metaphysics in concepts or language. Yet these are sectarian views about the proper scope and ambitions of philosophy. Many philosophers, perhaps the majority, believe philosophy can aspire to do more. Among those other possible theoretical ambitions is the supplementing or overturning of common sense, reconfiguring received wisdom, or ushering in gestalt-shifts in our conception of things, whether for wisdom or truth.

So, we have two sets of cross-cutting distinctions: (1) between first- and second-order theorizing and (2) between descriptive and prescriptive theorizing. Descriptive accounts of some target notion concern existing thought, talk, or practices. These distinctions are crosscutting in that we might have both descriptive and prescriptive theories at the level of first-order theorizing, and descriptive and prescriptive theories at the level of second-order theorizing. The way we think about a target notion might not be what we ought to think, and the principles by which we settle on the best first-order theory might not be the ones we should have.

Here is one way to represent those possibilities:

	diagnosis	prescription
first order	*a*	*b*
second order	*c*	*d*

So, a theory about some target notion can be a theory about either (*a*) ordinary convictions about some subject matter, or (*b*) what we ought to believe about that thing. Yet,

we often arrive at first-order theories based on background considerations about what phenomena, convictions, or principles need to be accommodated within a theory—that is, because of second-order convictions. At the second order, the difference between how we think (*c*) and how we ought to think (*d*) arises again. That is, there is a difference between what second-order principles we have for deciding first order questions (for example, 'cohere with this religion!' or 'vindicate our commonsense convictions!') and what second-order principles we ought to accept for settling first-order commitments. In turn, those second-order commitments are subject to yet higher-order disagreements, including third- or even higher-order principles concerning the principles for preferring lower-order principles.

The history of our understanding of a variety of subject matters has worked through transitions between the various cells on the above diagram, although perhaps not in any standard order. A toy model of the history of physics illustrates one way this can go. Our first physical theories were what we might now call "folk physics"—mostly naïve commonsense views about how the physical world works (cell *a*). This was presumably backed by a set of second-order theories, for example, that we should defer to our senses, or the local traditions, or what have you (cell *c*). Aristotle came along and thought that deferral to received wisdom was too quick. What we needed was to identify the essences and ends of things (cell *d*). While we could not entirely throw away our everyday understanding of physics, a good theory could depart from and even conflict with folk physics so long as it was rooted in Aristotle's preferred theoretical principles. This yielded a prescriptive theory of physics (cell *b*) that constituted a revision away from the comparatively naïve physics of Aristotle's predecessors.

Similar things might be said of subsequent revolutions in physics, including the Newtonian and Einsteinian ones. And of course, the basic phenomenon is not restricted to the sciences. Many of our theories start from received views, but modify their contents in light of conceptual innovations and ongoing theoretical pressures. A crucial advance in many domains has been the recognition that a good theory of some phenomenon need not match our folk theories about it. This sort of finding suggests we do well to accept a third-order principle of favoring second-order principles that tolerate departures from folk convictions.

The foregoing suggests some useful labels. At any order, a conventional theory offers a prescriptive theory that is consistent with one's diagnostic account at that order. A conventional theory can be more detailed or elaborated, with commitments that exceed the commitments contained in one's diagnostic account. For example, a conventional theory of, say, perception might appeal to neurological processes of which most of us are unaware. In contrast, a theory is revisionary if it has a prescriptive account that conflicts with elements of its diagnostic account. Because conflicts can be a matter of degree, the extent to which a theory is revisionary is also a matter of degree. In all these cases, the question of whether a

theory is revisionary is indexed to some account of what the received views are—whether ordinary persons or theorists.[1]

With these distinctions in mind, let us return to the topic of free will.

The Free Will Debates

When it comes to options for first-order theories about free will, we have an embarrassment of riches. There are libertarian accounts of various stripes, including agent- and event-causalists. Compatibilist accounts include "mesh" or self-expression theories, reasons-responsiveness theories, and so on. There are also different versions of free will skepticism and eliminativism, including impossibilists who hold that free will requires something impossible, and others who hold that it is possible but that we simply lack it. There are also various minority positions that have enjoyed different degrees of support at different times—classical conditional analysis compatibilists, uncaused event libertarians, two-standpoint views, proponents of Kantian noumenal freedom, and many more besides.

Yet, even if we had a maximally accurate theory of free will, it is unclear that we would recognize it as such. There is no obvious metric to recognize its superiority to alternative accounts. There are some of the usual theoretical virtues we might appeal to—parsimony, explanatory power, consilience, and so on. Still, any measure of these things seems downstream from some account of what a theory of free will has to show.

The fight over whether free will is compatible with determinism—arguably the central dispute in the philosophical literature over the past century—is symptomatic of the underlying issue. Proponents of the major positions attempt to motivate their views on the compatibility debate by appealing to radically different considerations. Some appeal to phenomenology and to the intuitiveness of a favored picture of agency or freedom. Others invoke the putative meaning of words to resolve the dispute. Still others dismiss all these

[1] For overviews of contemporary revisionist theories and the details of their relationship to conventional theories, see McCormick (2016), Vargas (2023), and Vargas (forthcoming). For a complementary account of similar distinctions as they pertain to metaethics, see McPherson and Plunkett (unpublished). For the account proposed in the present article, 'revisionism' and its cognates are used to refer to theories that emphasize conceptual revision, as opposed to theories that retain our concept but argue for revisions in how we apply the concept. For example, the conceptual revisionist about free will, gender, or marriage argues for a new understanding of free will, gender, or marriage that conflicts with received views about these things. A person who accepts traditional understandings of free will, gender, or marriage might insist on retaining those concepts in their traditional forms, while advocating that we change our assessment of whether animals have free will, whether it is stylish to sport a beard, or whether marriage can be performed outside of a religious building. Eliminativism or skepticism about some subject matter is usually best understood as offering a conventional account about the concept but radical revision or even elimination about the practices that are taken to depend on that concept.

things, and instead appeal to the apparent fixity of our psychologies, or the value of our practices. This is not to suggest that no one is deploying arguments, thought experiments, and the like. The contributors to these debates clearly are making arguments, and these arguments illuminate how different packages of commitments stack up. The point, though, is that all these arguments are embedded in diverse views about what the proper stakes are for debates about free will (Double 1996; Vargas 2013; Nichols 2015; Deery 2021; McCormick 2022).

This is not to say that there are no points of convergence among theorists. On perhaps the most common (and granular) way of fixing the subject matter of accounts of free will, free will is a kind of ability, power, or configuration of mental life that is required for the truth of, or ordinarily entailed by, direct attributions of moral responsibility.[2]

I am inclined to accept this characterization, but one might worry that this construal of free will seems to moralize it even though there are cases where, intuitively, one exercises free will without the stakes being moral. If I am deciding between a beautiful but less warm jacket or a dowdy but very functional coat, I seem to exercise free will despite the absence of morally significant stakes. The apparent ubiquity of non-moral exercises of free will raises the possibility that moralizing analyses may distort our accounts by drawing our attention away from potentially illustrative cases that are not themselves moral. That is a legitimate worry. Still, there are several things that can be said in defense of what we might call the responsibility-centric characterization of free will.

First, something can be a condition on some further thing while also being independently characterizable and important for yet other things. Having a working brain might be necessary for making choices about what to order for dinner at a restaurant, but that does not mean that what a working brain *is* is a dinner-ordering organ. That free will is characterized in terms of its centrality for moral responsibility does require that this be all there is to free will. This characterization helps us isolate a role or function—perhaps one of many—that free will needs to satisfy for us to recognize it as the thing we are trying to talk about.

Second, and relatedly, one might think that what is at stake here is a kind of agency that makes one a candidate for responsibility—that is, a responsible agent. Yet one can be an

---

[2] Why "direct," that is, non-derivative attributions? There might be cases where one has knowingly and intentionally done something to impair or preclude one's having free will, for example, rendering oneself unconscious or with a radically altered consciousness, or having put oneself in a coercive or forced-choice situation. In those cases, some have thought one can get culpability for actions done under those conditions, but that culpability is derivative of the upstream choice to constrain one's freedom. The issues here get tricky when we consider what had to be knowable in advance. For present purposes, we can bracket these issues but see Vargas (2005a) and Fischer and Tognazzini (2009) for discussion.

apt candidate for responsibility, understood as moral culpability and credit, without thinking that one's culpability and credit are settled solely by being a responsible agent. Blameworthiness for some bit of behavior is not determined solely by general facts about what kind of agent you are. It also depends on a variety of further factors—for example, whether you acted wrongly, whether it is fair to hold you responsible, or whether you had suitable opportunities for non-culpable action (Wallace 1994; Fischer and Ravizza 1998; Nelkin and Vargas forthcoming).

Third, a responsibility-centric conception need not be restricted to moral responsibility. After all, we apportion blame and credit in virtually any context where controlled adherence to norms is of some significance to us. Athletes and artists alike can be praised for innovation and condemned for mistakes. When an athlete tries to embarrass his teammates by playing poorly, or when an artist makes art to further the immoral regime, such things plausibly have moral valence. Yet, even a moral responsibility skeptic might insist that it reflects badly on the artist's aesthetic credentials that the sculpture is mediocre, that the kicker's kick was off-frame, or that the mathematician's proof was inelegant. Where there are norms—moral or otherwise—we seem to have some notion of blame and credit for meeting those norms.[3]

Even if we accept a responsibility-centric conception of free will, one might be skeptical that this does much work in fixing the topic. After all, philosophers have noted a variety of responsibility "concepts" "senses," "faces," or "aspects" (Watson 1996; Vincent 2011; Fischer and Tognazzini 2011; Shoemaker 2011). In linking free will to responsibility, one might object that we have substituted one philosophical quagmire for another. To avoid this result, we would need some way of specifying the kind of responsibility that is taken to fix the subject matter of debates about free will.

Interlude

Before concluding with an alternative picture of the subject matter of free will, it may be instructive to reconsider the constraints facing any satisfactory effort to establish that subject matter.

Recall the distinctions I introduced at the outset, between diagnostic and prescriptive theorizing, and between conventional and revisionary theories. The conventional theorist holds that a satisfactory theory needs to cohere with the contents of our diagnostic theory. Crudely: it is a cost to a philosophical theory if it conflicts with intuitive, received views about that thing. In contrast, revisionary theories hold that there

[3] Elsewhere, I distinguish between narrower and wider notions of free will, with the narrower emphasizing sensitivity to specifically moral considerations, and the broader notion emphasizing any normative notions (Vargas forthcoming).

can be good reasons for a prescriptive account of something—marriage, water, free will—to conflict with elements of some received view about the nature of that thing.

As we saw, there are a variety of reasons to allow for the possibility of revisionary theories in our formulation of second-order principles for specifying the subject matter of first-order disagreements. After all, sometimes we have false beliefs about the nature of things, and those false beliefs can function as impediments to a better theory. Revisionary theories have constituted much of the history of science, and really, our expanding understanding of the world. The point is that there are different ways to think about a theory's ambitions, and although one can prefer conventional theorizing, this should not lead us to stack the deck against the possibility of revisionism in a given domain.

However, this thought might seem like a nonstarter if you doubt whether there is some fact of the matter about what responsibility (or free will) is like, apart from our concept of it. Michael McKenna (2009) articulates this thought in the context of an earlier and related debate:

> Consider one of the comparisons Vargas makes between the free will issue and the ordinary concept of water. In the latter case, we discovered something about water, and we revised our concept of it accordingly. But the problem here is that water is something we can rigidly designate. We can be a realist about exemplar cases of it, and then we can ask whether its real nature is as we understand it to be. . . Neither free will nor moral responsibility seem to be like that. Think about it from Pereboom's perspective, that is, from the perspective of a free will and moral responsibility anti-realist. If there is no thing that it is to be morally responsible (in the sense Pereboom is interested in), at least at this world, then what moral responsibility is cannot come apart from the concept in such a way that there is, so to speak, something for moral responsibility to be beyond our concept of it (10-11).

McKenna is articulating at least three ideas. One is that revisionism presumes realism. Let us put that to one side.[4] The second thought is McKenna's endorsement of the idea that Pereboom does not seem to think a revisionist possibility is available here, and the third is McKenna's conjecture about why that seems so. McKenna interprets Pereboom as holding that free will and moral responsibility are not like water, and that the appeal of revisionism

[4] I am unpersuaded that a revisionary picture presupposes realism about that thing, where realism means that one is committed to its existence and that existence is at least partly independent of subjective beliefs. A revisionary theory about some target notion requires that one think that thought and talk about that thing admits of truth and falsity. But that is short of a commitment to realism. For example, one might think the best theory of true love conflicts with various folk beliefs about love (such a theory would be revisionist), but that true love nevertheless does not exist (thus making the theory anti-realist) (Vargas 2023). See also Vargas (forthcoming) for a related discussion about "patchy realism."

depends on the distinctive nature of water. Revisionism, he seems to think, requires some features that water has but that responsibility does not.

I am unpersuaded that things stack up in this way. First, many people are realists about free will and moral responsibility (as McKenna acknowledges, he is one of them). Second, revisionism doesn't depend on natural kind terms, and it does not obviously depend on rigid designation. As we saw at the outset, there are revisionary accounts of social, moral, and artifactual kinds, including race, virtue, marriage, and so on. To be sure, in each of these cases we need some explanation about why the revision is not a topic change. But the point is that revisionary theorizing is not limited to scientific kinds like water. So, although I am inclined to accept McKenna's diagnosis of Pereboom's presumptions, I think we have reason to reject the idea that revisionism about free will and moral responsibility is obviously a nonstarter.

There is something plausible about McKenna's second and third ideas: Pereboom seems to suppose that we must be conventionalists at the first and second orders, for there is no way to theorize absent the folk concept. He seems to think that a theory of free will has to be constrained by our views about the nature of free will.[5] If we have reason to doubt the existence of free will characterized by our received views about its nature, then he thinks we should be hard incompatibilists. This is, of course, itself a higher-order view about how to specify the subject matter of first-order debates.

We can, however, re-ask the prescriptive question—are these principles (principles that take our received picture of free will as sacrosanct) ones that, all things considered, we ought to accept? I am inclined to think not. Again, I think we can and oftentimes do have false beliefs about the nature of things. The important thing to keep track of here is, as Pereboom (2021, 11) and others have noted, that it is an important desideratum for higher-order proposals that they capture, explain, or otherwise allow us to distinguish the partisans of the lower-order debate. So, whatever (second-order) view we adopt about the proper subject matter of first-order debates about free will, it needs to allow us to capture the possibility of substantive disagreements at the second-level.[6]

---

[5] McKenna is surely right to characterize Pereboom as committed to what I have here termed conventionalist theorizing about the nature of free will. This is evident both from the general structure of eliminativisms, which tend to rely on arguments from the non-existence of some thing understood in a non-revisionary way (Vargas forthcoming) and in the particular details of Pereboom's various replies to conceptually (as opposed to practice-based) revisionary projects (for examples, see Pereboom 2007, 200-201; McKenna and Pereboom 2016, 291-292).

[6] Here, one might worry that infinite regresses are possible. Perhaps they are, but I do not think that is the issue. What we need is sufficient convergence at some higher level (second, third, or nth) that allows us to make sense of the apparent fact of there being a substantive debate about free will. If it turns out that there is no

A promising approach

Our challenge is to vindicate the responsibility-centric picture of free will by fixing our target without heavily relying on our potentially error-ridden convictions about the nature of responsibility. To do that, we need some way of specifying what phenomena to look for, one that does not require an articulation of conceptually essential features. Here is an alternative: we can fix the target phenomena by appeal to some truisms about responsibility (Vargas, forthcoming). The notion of responsibility at stake are the ones in everyday life, whose recognizable features include blame and credit, excuses, some idea of desert, and so on. These generalities are just that; they are not robust enough in content to stipulate anything as precise as Pereboom's basic-desert sense of responsibility. Fortunately, they do not need to be. All they must do is to point us to some functional roles or relations that help us fix the phenomena in the world that are the subject of our theorizing. That is the sense in which the approach is a broadly functionalist one.

(Notice that the general strategy is continuous with functionalist theorizing in many domains. The general functionalist strategy across many philosophical contexts is to make progress on a contested issue by focusing on some target phenomena, and asking what realizes them, explains them, or could justify them, without requiring that this goes through some armchair-specified account of hidden essences. This is compatible with the possibility of some target phenomenon having multiple functions and multiple nested functions within some larger functional system or set of systems. It is always a potential discovery, to be evaluated in the usual ways, that some phenomenon has multiple functions.)

The relatively general truisms gestured at above are sufficient to give us a reasonably clear set of target phenomena. We're looking at everyday responsibility practices: the finding of fault, the business of apportioning credit and blame, of offering excuses, and the like. That is the kind of responsibility at stake in theories of free will. What that comes to—including what is required to justify practices like these—is a matter to be discovered, not stipulated. On this picture, first-order theories of free will are proposals about the kind of freedom, control, or agency that is required to justify our engaging in these kinds of practices.

---

convergence at any higher level to support some fixing of the proper domain of first-order disagreements, then we have discovered that, appearances aside, there is not enough background agreement for first order disagreements to be more than verbal disputes. However, if there is enough higher-order agreement about the subject matter of the debate, or how to determine the subject matter of the debate (or about how to determine how we determine the subject matter of the debate . . . and so on), then we have some reason to think it is not entirely hopeless that we can get some convergence around an account of what the free will debate is (properly) about.

Recall the prior example of marriage. The theorist who starts from conceptually identified essences might try to construct a rich theory of marriage by considering everyday convictions about the nature of marriage. If that theorist lives in a religious community, they might insist that the essence of marriage is a sacramental relation between a man and a woman (and perhaps a divine being). Such a theorist might make good on local intuitions about marriage, but the resultant theory would do a poor job of capturing the wide diversity of practices that have been recognized as forms of marriage more generally.

In contrast, on the functionalist approach suggested here, we put aside the question of what the conceptual essence of marriage comes to and instead appeal to some truisms about marriage (for example, that it is a privileged social relation, usually understood to involve the linking of family units and family formation, distinctive social privileges regarding speaking for spouses, and so on). In articulating truisms, we are less concerned with getting the conceptual contours exactly right than we are interested in triangulating some phenomena of interest in the world. Accordingly, once we have our stockpile of truisms, we then look at the phenomena of the world, and see what sorts of practices realize these things. Were we further interested in arguments about whether marriage is good, bad, or otherwise, we could ask questions about whether those practices are justifiable, have good effects, and so on.

This is not the place for a more systematic defense of this kind of approach to the free will problem, although I have made efforts in that vein elsewhere (Vargas 2022; forthcoming). Here, my main ambition has been to highlight the general structure of important disagreements in debates about free will and to call attention to some shortcomings of a widespread methodology that privileges a putative conceptual core for theorizing about free will. Given that there is a promising alternative method of theory construction, one that does not presuppose that our introspective efforts are reliable guides to the world, we have reason to think the proper subject matter of free will debates does not need to depend on capturing some armchair specification of a privileged concept of responsibility, desert, or free will.

I conclude with a pair of observations about the power of this way of characterizing the nature of first-order debates about free will.

First, the proposal respects the possibility we noted at the outset, that a good theory of something might depart from received views about that thing. Moreover, the picture allows that this could be a matter of degrees. It might turn out that some but not all parts of our extant responsibility practices can be justified. Or it might turn out that everything must go, or alternately, that everything is in good standing. Even so, nothing in this picture requires that we be revisionists; it is entirely compatible with this construal of the stakes that the best theory is a conventional one, where the first-order prescriptive element neatly

coheres with the first-order diagnostic account of our received views about free will and moral responsibility. Functionalism does not commit us to revisionism about free will or anything else, but it rightly permits that possibility.

Second, the approach respects the injunction to make sense of the main disagreements (Pereboom 2021). The core issue on this picture of the stakes is which notion or notions of free will can justify and make sense of the kinds of responsibility practices we have. Incompatibilists have their preferred answers, and compatibilists have theirs. Conventionalists have theirs, as do revisionists. These are all live possibilities, and so too is eliminativism. For all that has been said here, it might turn out that we cannot identify any picture of freedom that is sufficient to license responsibility practices like ours.

References

Deery, Oisín. 2021. Naturally Free Action. Oxford: OUP.

Dennett, Daniel. 1984. Elbow Room. Cambridge: MIT.

—. 2003. Freedom Evolves. New York: Viking.

Double, Richard. 1996. Metaphilosophy and Free Will. Oxford: OUP.

Fischer, John Martin, and Mark Ravizza. 1998. Responsibility and Control: A Theory of Moral Responsibility. NY: Cambridge UP.

Fischer, John Martin, Robert Kane, Derk Pereboom, and Manuel Vargas. 2007. Four Views on Free Will. Malden, MA: Wiley-Blackwell.

Fischer, John Martin, and Neal Tognazzini. 2009. "The Truth About Tracing." Nous 43 (3): 531–56.

—. 2011. "The Physiognomy of Responsibility." Philosophy and Phenomenological Research 82 (2): 381–417.

McCormick, Kelly. 2016. "Revisionism." In Routledge Companion to Free Will, edited by Kevin Timpe, Meghan Griffith, and Neil Levy, 109–20. New York: Routledge.

—. 2022. The Problem of Blame: Making Sense of Moral Anger. Cambridge: Cambridge UP.

McKenna, Michael. 2009. "Compatibilism and Desert: Critical Comments on Four Views on Free Will." Philosophical Studies 144 3–13.

McKenna, Michael, and Derk Pereboom. 2016. Free Will: A Contemporary Introduction. New York: Routledge.

McPherson, Tristram and David Plunkett. (m.s.) "After Metaethics." Unpublished manuscript.

Nelkin, Dana Kay, and Manuel Vargas. Forthcoming. "Responsibility and Reasons-Responsiveness." In Freedom, Responsibility, and Value: Essays in Honor of John Martin Fischer, edited by Taylor Cyr, Andrew Law, and Neal A. Tognazzini, New York: Routledge.

Nichols, Shaun. 2015. Bound: Essays on Free Will and Responsibility. Oxford: OUP.

Pereboom, Derk. 2007. "Response to Kane, Fischer, and Vargas." In Four Views on Free Will, Malden, MA: Wiley-Blackwell.

—. 2021. Wrongdoing and the Moral Emotions. Oxford: OUP.

Smart, J.J.C. 1961. "Free Will, Praise, and Blame." Mind 70 291–306.

Strawson, P. F. (1962). Freedom and Resentment. Proceedings of the British Academy, XLVIII, 1-25.

Vargas, Manuel. 2005a. "The Trouble with Tracing." Midwest Studies in Philosophy 29 (1): 269–91.

—. 2005. "Compatibilism Evolves? On Some Varieties of Dennett Worth Wanting." Metaphilosophy 36 (4): 460–75.
—. 2013. Building Better Beings: A Theory of Moral Responsibility. Oxford: OUP.
—. 2015. "Desert, Responsibility, and Justification: Reply to Doris, McGeer, and Robinson." Philosophical Studies 172 (10): 2659–78.
—. 2022. "Instrumentalist Theories of Moral Responsibility." In The Oxford Handbook of Moral Responsibility, edited by Dana Nelkin, and Derk Pereboom, 3–26.
—. 2023. "Revisionism." In A Companion to Free Will, edited by Joseph Campbell, Kristin M. Mickelson, and V. Alan White, 204–20. Oxford: Wiley-Blackwell.
—. Forthcoming. "Revisionism." In Four Views on Free Will: Second Edition. Malden, MA: Wiley-Blackwell.
Vincent, Nicole. 2011. "A Structured Taxonomy of Responsibility Concepts." In Moral Responsibility: Beyond Free Will and Determinism, edited by Nicole Vincent, Ibo van de Poel, and Jeroen van den Hoven, 15–35. Dordrecht, The Netherlands: Springer.
Watson, Gary. 1996. "Two Faces of Responsibility." Philosophical Topics 24 227–48.