A Discretionary Case for Preservationism about Free Will

Kelly McCormick *
k.mccormick@tcu.edu

ABSTRACT

How does the term 'free will' refer? This question seems to lie at the center of debates about whether the attitudes and practices that depend on our successful attributions of basic-desert-entailing moral responsibility ought to be preserved or eliminated. In this paper I tackle questions about the way that different reference-fixing conventions might inform disagreement between preservationists and eliminativists about free will and moral responsibility, and argue that even recent elimination-friendly work on reference fails to offer much real support for eliminativism. In fact, making explicit the role that different *motivating concerns* play in rendering certain reference-fixing conventions operative for eliminativists and preservationists suggests at least one powerful reference-based argument in favor of preservationism.

1. Introduction

It is often the case that especially well-worn and seemingly intractable philosophical debates come to enjoy a dialectical second-wind – an infusion of contemporary interest and energy – by way of a sort of s*keptical renaissance*. The pattern here will be familiar from a variety of areas in which apparent dialectical stalemates about the nature of some feature of the world come to be radically recast by the proposal that perhaps the kind of thing in question doesn't really *exist* at all. Consider, for example, Unger (1979) and van Inwagen (1990) on composite objects[1], Churchland (1981, 1986) and Stitch (1983) on common sense folk-psychological concepts like belief, Feyeraband (1962) and Laudan (1984) on scientific realism, Griffiths (1979) on emotion, Machery (2009) on concepts, Mackie (1977) and Blackburn (1985) on moral properties, and Appiah (1995), Andreasen (2000), and Mallon (2006) on race. For debates that

* Texas Christian University, USA.

[1] Though van Inwagen allows for the existence of composite living organisms.

have undergone this kind of skeptical renaissance the central point of conflict in the dialectic often shifts to disagreement between *preservationists* and *eliminativists* about the kind of thing in question. Preservationists – sometimes retaining and defending a traditional view and sometimes proposing and defending a substantially *revisionary*[2] one – are those who maintain that the kind of thing in question exists. Tables and chairs, beliefs, and moral properties are robust features of the world, even if they turn out to be somewhat – or even radically – different from the way that we currently think about them. Eliminativists disagree.

But what is the nature of this disagreement? While I have discussed it at length elsewhere[3] here I will focus on one characterization in particular that I think merits closer attention, namely as a disagreement about whether or not the term 'free will' ever successfully refers.

For much of the twentieth century assumptions about reference were rarely made explicit in debates about free will and moral responsibility, but there have been a handful of noteworthy exceptions.[4] First, Mark Heller (1996) offered early insights about the way that different reference-fixing conventions might inform how we assess traditional compatibilist views about free will. More recently, Shaun Nichols (2015) has systematically explored the way that our reference-fixing conventions shape debates about the existence of free will, ultimately proposing the *discretionary view* that 'free will' is an ambiguous kind term. According to this novel form of pluralism preservationists and eliminativists both speak truly when they make claims about the existence of free will. Each appropriately deploys a different reference-fixing convention for the term 'free will', one (for preservationists) that allows the term to successfully refer and another (for eliminativists) that does not. However, Gregg Caruso (2015) has recently claimed that Nichols' discretionary view can in fact be used to motivate full-blown eliminativism.

Here I will argue that we do well to attend more carefully to the role that different reference-fixing conventions play in debates between preservationists and eliminativists about free will, but that both Nichols and Caruso reach the wrong conclusions. Rather than recommending eliminativism or even pluralism,

[2] See Vargas (2013) for the most comprehensive development of a revisionary preservationist account of free will and moral responsibility currently on offer.

[3] See McCormick (2019, 2022).

[4] In addition to those mentioned here, see also Hurley (2000), Vargas (2017), and Deery (2021a, 2021b).

a discretionary view about the possible operative reference-fixing conventions for 'free will' in fact suggests good reason to embrace preservationism.[5]

I begin in Section 1 with some brief clarificatory remarks on the distinction between causal historical and descriptive reference-fixing conventions. In Section 2 I turn to discussion of Nichols' pluralist discretionary view of free will, as well as Caruso's argument that it can be used to motivate full-blown eliminativism. In Section 3 I argue that Caruso is subject to a dilemma, and that the discretionary view is not in fact well suited to motivate eliminativism. However, in Section 4 I argue that explicit attention to the way that different reference-fixing conventions can impact the truth of existence claims about free will highlights the central role that *motivating concerns* play in the debate between preservationists and eliminativists. While eliminativists are often motivated by concerns about the permissibility of harming those who do not deserve it, preservationists are often motivated by more victim-centered concerns. And, while the pressing nature of *both* these kinds of concerns might go some way toward ultimately explaining the apparently intractable feel of the disagreement between preservationists and eliminativists, here I will throw my hat in with preservationists and argue that we ought to prioritize their victim-centered concerns.

2. Two Possible Reference-fixing Conventions for 'Free Will'

Mark Heller (1996) is one of the first to suggest that influential twentieth century insights about reference might fruitfully be extended to debates about free will. Heller identifies and pushes back on implicit assumptions about the way that 'free will' refers suggested by traditional theorizing about free will via conceptual analysis or as a Lewisian theoretical term.[6] According to Lewis, for example, theoretical terms get their meaning and reference from the theory in

[5] Here it is worth noting that Oisín Deery (2021b) offers what I take to be a powerful parallel argument that Nichols' discretionism in fact recommends preservation over pluralism, and that Caruso's arguments for eliminativism can be overcome. However, the success of Deery's argument requires taking on board certain commitments regarding his own view that free will is a genuine kind, that the phenomenology of free will is not only libertarian but also *accurately* so, and explicit realism about free will. While I find Deery's arguments for each of these commitments both carefully developed and downright persuasive, my goal here is to offer a parallel line of argument for preservationism which avoids at least some of the heavy – and to some costly – lifting of Deery's overall theoretical package.

[6] See Lewis (1972).

which they are embedded. The term in question refers to whatever feature of the world satisfies some crucial set of claims in the theory. If nothing satisfies these claims, then reference fails. On this kind of view reference is often tied closely to our concept of the thing in question, and successful reference often does not permit significant error when it comes to our platitudinous beliefs, folk concept, or best conceptual analysis of the thing in question. Here I will refer to this kind of conservative[7], conceptually anchored reference-fixing convention generally as a *descriptive convention*.

Heller's key insight is that instead of defaulting to a descriptive convention we might think of 'free will' as a *kind term*, and that doing so raises the possibility of distinguishing between the extension of this term and the concept associated with it.[8] As Putnam (1962) and Kripke (1980) famously noted, the two might come apart in interesting ways. Putnam (1962, 1975), for example, suggests that we might come to *discover* facts about kinds that are absent from – or even at odds with – our concepts of them. The essential properties of a kind might not fit our concept, but this does not entail that the associated kind term fails to refer.

In order to elucidate this point Heller cites Putnam's famous robot cat example. If we came to discover that the fluffy companions we have been calling cats turned out to be automata controlled by Martians, this does not mean we should conclude that there are no cats. Of course, if our *concept* fixed reference then we would have to conclude that the extension of 'cat' turns out to be empty.[9] But Putnam argues that there is another possibility. Perhaps instead the extension is determined by *paradigm cases*. Cats are just anything that is of the *same kind* as whatever the paradigm cases turn out to be. One such paradigm is staring at me now as I type this. If it turns out that this fluffy thing I've been calling a cat (his proper name is Butters) has been a robot controlled by Martians all along, then insofar as the other paradigms (for example, another fluffy thing that I have named Coco) turn out to be of the same kind, then Putnam's recommendation is that we ought to revise our concept of cats rather than say that it turns out that

[7] In the sense that it sets the bar for a theoretical term to successfully refer rather high.

[8] For further development of the proposal that we take free will to be a genuine kind see Deery (2021b).

[9] At least according to Putnam's usage of the term 'concept'. Given lack of consensus on what concepts themselves amount to it is perhaps worth noting that drastically different views of concepts might allow for conceptually fixed reference without entailing that the extension of 'cat' turns out to be empty in this case (see, for example, Fodor (1998)).

there are no cats. For Putnam, the essential nature of a genuine kind is an empirical matter, something to be discovered, and like many discoveries the results might be (and often are) surprising.

Heller's own suggestion is that we might extend this line of reasoning to free will. It is at least prima facie plausible that free will is a genuine kind. If so, and Putnam's insight on how kind terms refer is also correct, then Heller suggests that assuming a descriptive view of how 'free will' refers and the standard method of subjecting views of free will to "death by counterexample" are both misguided (Heller, 1996: 334). The standard method presupposes that conceptual analysis is the appropriate methodology for theorizing about free will, and that it is ultimately our concept of free will that fixes its reference and thus determines its extension. But, once we consider free will as a kind this method is no longer obviously appropriate. Instead, we ought to take *paradigms of free action* as our starting point, and work to discover what the essential nature of those paradigms is.[10] Combined with a story about initial baptism (when the term 'free will' is first introduced) followed by the right kind of causal-historical chain of transmission between speakers from there, we get an alternative account of reference-fixing that I will hereafter refer to as a *causal historical convention*.

Here I do not intend the contrast between descriptive and causal historical conventions to exhaust all of the possible ways that a term like 'free will' might refer, but they do largely capture the variety of conventions discussed explicitly by those interested in adjudicating between preservationists and eliminativists about free will. I turn now to Shaun Nichols' novel suggestion that these two distinct reference-fixing conventions may not be mutually exclusive, but rather one might be operative for preservationists and the other for eliminativists *at the same time*.

3. Discretionism, Pluralism, and Eliminativism

As initially formulated by Nichols (2015) the discretionary view delivers a kind of pluralism about the truth of existence claims about free will. In some contexts

[10] Heller's own proposal is that these insights might be used to defend purportedly counterintuitive varieties of compatibilism in particular. For further discussion of Heller's specific argument, and objections to it, see Daw and Alter (2001). And, for extended discussion of Heller's proposal, Daw and Alter's objection, and suggested problems with the latter see Deery (2021b), Chapter 2.

reference succeeds, but in others it does not, and so according to Nichols eliminativists and preservationists can both speak truly without contradiction. However, Gregg Caruso (2015) has recently attempted to make use of Nichols' discretionary view to motivate full-blown eliminativism rather than pluralism. Here I discuss each of these views in turn.

3.1 Nichols' Discretionary View

Nichols first tackles a variety of descriptive questions about our folk concepts of agency, determinism, and moral responsibility in the service of providing a folk psychological diagnosis of the problem of free will. He argues that this problem stems from fundamental conflicting intuitions about the nature of agency: that starting from childhood we are compulsive seekers of deterministic causal explanation, though at the same time find it deeply counterintuitive to think of our own decisions as determined. Further, the belief that our decisions are indeterministic is unjustified, because it rests on the faulty assumption that *if* these decisions were determined then we would *know* that they were. All of this, Nichols argues, suggests a debunking argument that can be used to supplement traditional arguments against libertarianism.

Should any of this lead us to believe that free will does not exist? Here Nichols offers a novel approach to assessing the disagreement between eliminativists and preservationists more broadly, and suggests that we adopt a *discretionary view* about who is correct. According to this view, the term 'free will' is an ambiguous kind term. For Nichols, this means that the referent of its tokens is fixed by different conventions in different circumstances. So, when eliminativists say, "Free will does not exist," and preservationists say, "Free will does exist (though it may not be quite what we thought it was)," it is possible that both speak truly. Something about the circumstances in which these two utterances occur could make different reference conventions operative for each.

But which features of the circumstances of utterance matter here? According to Nichols, the operative reference-fixing convention for a token of the term 'free will' will depend in some way on our *practical interests*:

> Although the actual historical role of practical interests is unclear, it is very plausible that practical interests can have important effects on ontological claims....it does seem likely that practical considerations can impact which [reference] conventions we adopt. In addition, if Pinillos, Mallon, and I are right about the availability of different reference conventions, then there need be no mistake in adopting one convention or the other (Nichols et al. 2016). As a

> result, we might appeal to practical interests in deciding which convention to adopt and impose. (Nichols, 2015: 69)

Importantly, on the discretionary view there could be sufficient differences in the practical interests relevant to the circumstances of eliminativists' and preservationists' utterances such that different reference-fixing conventions are operative for each. For example, a descriptive convention may be operative for eliminativists because something about their interest in free will renders the more conservative convention appropriate. When eliminativists says, "Free will does not exist," the extension of 'free will' is picked out by a description of a kind of agency that has no application (perhaps ultimate sourcehood, or being *causa sui*). Thus, like 'phlogiston', for eliminativists the extension of 'free will' is empty, and what they say is true. In contrast, the interests of preservationists might render a more liberal causal-historical convention operative. When preservationists say, "Free will exists," the extension of 'free will' is picked out by the intentions of a speaker in an initial baptism along with the similar intentions of subsequent speakers who make paradigmatic attributions of 'free will'. When this convention is operative, it is charitable to assume that the speaker intends to pick out an existing form of agency (perhaps reasons-responsiveness, or the right kind of identification between one's real self and their action), even if they have some mistaken beliefs about it. For preservationists, then, the extension of 'free will' is not empty and what they say is also true.

On this view the referential ambiguity of 'free will' allows for the fact that both a descriptive convention and a causal-historical convention are *available*, but which convention is actually *operative* for a given speaker will in some way depend on the speaker's reasons for being interested in free will in the first place. If this is the case, and preservationists and eliminativists have very different primary *motivating concerns*, then again different reference-fixing conventions might be operative for each.[11] Nichols himself demonstrates the way that this kind of pluralism might go by using Galen Strawson's (1994) eliminativist impossibilism and Manuel Vargas's (2013) preservationist revisionism as instructive examples:

> Now we can finally get to the issue concerning eliminativism and preservationism about free will. We have been assuming that the folk conception of free will contains significant error. If the foregoing story about the diversity of reference conventions is right, how should we interpret Galen Strawson when he says "Free

[11] I will say a great deal more about the role of motivating concerns in Section 4 below.

> will doesn't exist"? Descriptively, of course. He is keying on the false description associated with "free will," and pointing out that nothing meets that description. To interpret Strawson's use of "free will" causal-historically would be manifestly uncharitable. What reference-convention is in place when Manuel Vargas says "Free will isn't what we thought"? Presumably *not* restrictive descriptivism, or what he says is, by his own lights, false. *This allows us to say that Vargas is right and Strawson is also right.* It's just that the term "free will" operates with a different reference convention in the different contexts. (Nichols, 2015: 66; emphasis my own)

Nichols (2015: 66) acknowledges that this pluralism "deflates somewhat the importance of the metaphysical dispute between eliminativists and preservationists," a feature of the discretionary view which has unsurprisingly drawn criticism from both eliminativists and preservationists alike.[12] I turn now to an eliminativist criticism recently offered by Gregg Caruso.

3.2 Caruso's Discretionary Case for Eliminativism

While Caruso (2015) pushes back on Nichols' pluralist conclusion his argument proceeds by first granting a number of Nichols' own assumptions:

> I am willing to grant for the sake of argument that (1) the concept of "free will" is enmeshed in significant error, (2) the free will debate depends on substantive assumptions about reference, (3) not all theoretical terms embedded in false theories should be eliminated, and (4) reference is systemically ambiguous. (Caruso, 2015: 2827)

Caruso also agrees with Nichols' suggestion that a descriptive reference-fixing convention is operative when eliminativists say, "Free will does not exist," and so they speak truly. Where he and Nichols diverge is in regard to the claim that a causal historical convention can ever render preservationists' claim that free will exists true as well. Caruso argues that even if this more liberal reference-fixing convention is operative for preservationists, we *still* have good reason to think that 'free will' fails to refer. And so, on Caruso's version of the discretionary view eliminativists speak truly when they claim that free will does not exist, but preservationists' claims that free will exists are false.

Why think that 'free will' fails to refer, even if a causal historical reference-fixing convention is operative for preservationists? First, while a causal historical convention tends to encourage quantification over actual features of

[12] See, for example, Vargas (2017) and Kane (2017). For Nichols' responses to Vargas and Kane, as well as some of the arguments here, see Nichols (2017).

the world, it does not *guarantee* successful reference. Whether or not a term successfully refers when a causal historical convention is operative will depend on at least three things: (1) facts about the initial baptism of the term, (2) facts about the causal chain between initial baptism and the present speaker, and (3) whether or not the paradigms in our contexts of use are sufficiently similar to take them as involving a genuine kind. While Caruso is primarily interested in arguing that (1) leads to reference failure, it may also be helpful to say a bit more about each of the other two possibilities. In regard to (2), successful reference will depend on whether or not the intentions of the speaker in the initial baptism plausibly connect up with current usage via a non-deviant causal chain.[13] So, in cases of causal deviance reference might still fail even when a casual historical convention is operative. In regard to (3), reference could still fail if it turns out that we have good reason to think that the paradigms fail to pick out a genuine kind. Perhaps, for example, we come to realize that the similarity between paradigms is, at best, massively disjunctive.

Rather than pursuing either of these two possible paths to reference failure Caruso focuses on (1) exclusively and argues that something has gone wrong with our initial baptism of the term 'free will'. 'Free will' fails to refer in much the same way that 'phlogiston' would have if we imagine that a causal historical reference-fixing convention had been operative for Johan Becher. Had Becher posited the existence of phlogiston by, for example, demonstratively pointing to a pile of rust and then to some smoke rising from a fire and calling both things 'phlogiston', then his attempt to pick out a unified feature of the world would have failed. Such an attempt to successfully fix reference would have been a swing and a miss, given that it turns out there is nothing that unifies the phenomenon Becher would be attempting to get at via ostension. At best, we would get something like, "that stuff released when wood burns or metal rusts," but of course there is no such substance. Becher would have been "swinging" at an arbitrary disjunction of two different reactions involving oxygen, and simply striking out at picking out a genuine kind.[14]

[13] For example, a gradual shift in speakers' intension over time allows for continued successful reference, even while tolerating significant revision. But, a sudden and drastic shift (for example, a new assertion that we ought to be talking about *this new stuff* rather than *that old stuff*) might in turn shift reference to the extent that we have changed the subject.

[14] Here I am tempted to say that not even this path to elimination works if we *genuinely* adopt a causal historical account of reference, and do not implicitly sneak in theory-drive considerations. I discuss these consideration further in Section 3, and if I am right all the worse for Caruso.

Returning to free will, Caruso notes (I think correctly) that there is a difficulty in imagining how the initial baptism for 'free will' might have gone, given important differences between free will and the *observable* kinds that are often paradigms for an operative causal historical reference-fixing convention. While initial baptisms for observable kinds (like water and cats) were likely demonstrative, it is not at all clear what the demonstrative target might be in the case of free will.[15] Despite this difficulty, Caruso suggests several candidate targets before settling on the one that he prefers:

> It is possible, for example, that the initial baptism [for 'free will'] was to whatever power or ability is required to justify ascriptions of desert-based moral responsibility, or to that feature of choice and action that justifies our reactive attitudes, or to a set of compatibilist-friendly capacities (e.g. reasons responsiveness). While I cannot adequately address all these possibilities here (although I will say something about them below), my proposal is that we should look elsewhere, i.e. to the *phenomenology of free agency*. (Caruso, 2015: 2828)

According to Caruso there are several reasons for thinking that the initial baptism for 'free will' targeted the phenomenology of free agency. First, our first-person experience of agency is "primitive and basic" (Caruso, 2015: 2828). To support this claim Caruso appeals to intuitions about possible worlds in which we *lack* any first-person experience of free agency. Even if the other reference-fixing candidates mentioned above (for example, compatibilist friendly capacities) were present in such worlds, Caruso suggests that without this first person experience it is prima facie plausible to think that the term 'free will' would never have been introduced. Second, Caruso cites the prominent role that appeals to the phenomenology of free agency have played historically in arguments for libertarianism, especially those that cite our feeling of freedom as providing some degree of evidence for the existence of libertarian free will. While Caruso himself doubts the ultimate plausibility of such arguments, here he cites only their undeniable intuitive appeal, both amongst philosophers and the folk more generally.[16] Finally, Caruso argues that in contrast to the phenomenology of free agency other candidates for initial baptism are historically anachronistic (Ca-

[15] For further discussion of this kind of concern see McKenna (2009).

[16] Here Caruso emphasizes the role that appeals to the phenomenology of free agency have played in many agent-causal libertarian views in particular, for example Campbell (1957), O'Conner (1995), and Taylor (1992).

ruso, 2015: 2830). While a minimal condition of agency like reasons-responsiveness might be an obvious necessary condition for any plausible contemporary account of free will, it seems far too narrowly focused to capture "in *totality* our pre-theoretical self-conception as agents" (Caruso, 2015: 2830). Any plausible candidate for fixing the reference of 'free will' with an initial baptism must accommodate the fact that our use of this term stretches back millennia. As such it ought to be compatible with our pre-scientific, pre-theoretical views of ourselves as agents.[17] While the phenomenology of free agency looks to be an especially plausible candidate on this dimension, Caruso suggests that other more theoretically sophisticated compatibilist-friendly candidates such as reasons-responsiveness are not.

Caruso concludes that the phenomenology of free agency is the best candidate for the demonstrative target of any plausible initial baptism of 'free will'. This widely shared, basic, intuitive, and historically central first-person experience looks like just the sort of thing our ancestors might have been trying to get at in introducing the term 'free will' in the first place. It's *that* feeling – the feeling that your action is *up to you* in a specific way.

But in what specific way? If Caruso is correct then by Nichols' own lights the discretionary view would not yield pluralism. Even when a causal historical reference-fixing convention is operative, if the initial baptism of 'free will' targets the phenomenology of free agency then we find ourselves in one of the rare instances of causal historical reference failure. Much like the hypothetical example of Becher above, our ancestors would have been swinging and missing at ostensively picking out some actual, unified feature of the world. As discussed above, Nichols himself goes to great lengths to argue that our first person phenomenology of free agency is *libertarian*,[18] and also that this phenomenological experience is *in error*. If the phenomenology of agency is libertarian, but this illusory first person experience is the demonstrative target intended to fix the reference of 'free will', then even on a causal historical reference-fixing convention the term 'free will' will fail to refer. And so, Caruso concludes, even by Nichols' own lights we should be moved by the discretionary view to embrace eliminativism rather than pluralism.

[17] Caruso (2015: 2830) discusses the need for the relevant reference fixing feature to be compatible with dualism, in particular.

[18] For further argument in support of this claim also see Deery et al. (2013).

4. Caruso's Dilemma

Caruso's proposal that our first-person experience of libertarian agency could be a plausible target for the initial baptism of 'free will' is not without its merits. Here I will grant his claims that this experience is basic, widespread, historically pervasive, and intuitive. However, in this section I will argue that his claims about initial baptism and reference failure are subject to a dilemma. Either Caruso must implicitly sneak in the kind of theoretical and conceptual content relevant only to a *descriptive* reference-fixing convention to make his claims about reference failure plausible, or he must somehow make a case for the controversial claim that baptism is a one-time-only affair.

4.1 Implicit Conceptual Content

In claiming that the target of our initial baptism is the phenomenology of specifically *libertarian* agency Caruso implicitly sneaks theoretical and conceptual content into a reference-fixing picture that explicitly denies it this role. As he himself notes, paradigmatic examples of an operative causal historical reference-fixing convention – namely for observable kinds like water or cats – usually involve a relatively straightforward initial baptism. We observe that certain features of the world are similar to an extent sufficient to motivate our picking out *that* kind of thing as 'x'. But free will is not an observable kind, or at least not obviously so. As such, Caruso seems to assume that our only options for initial baptism will require appeal to *some* conceptual content in order to do the relevant reference-fixing work. And, borrowing from Nichols' (2015) own empirical work we should accept that the relevant content is specifically libertarian.

However, this is a mistake. Nichols' work concerns the content of our *concept* of free agency. But if a casual historical reference-fixing convention is truly operative for 'free will', then there is no clear place for such content in determining reference, and certainly not at the stage of initial baptism. This is precisely the point in distinguishing between causal historical and descriptive reference-fixing conventions in the first place. If it turns out that we cannot make sense of the initial baptism for a term without appeal to such content, this does not mean that we should simply sneak in *just a bit* in order to identify the kind of thing that we are talking about. Rather, if a casual historical convention is genuinely operative then such a move would never be necessary – some degree of similarity between paradigms is already sufficient. If it turns out that we need to appeal to conceptual content in order to identify this *similarity* – for example,

by aiming at specifically libertarian free agency – then we have a context in which a causal historical reference-fixing convention cannot properly get off the ground in the first place. This convention is simply not appropriate for the relevant term, and any conclusions we might draw about reference success or failure should instead proceed holding fixed the assumption that a descriptive convention is operative.

This conclusion by itself would, of course, only serve to strengthen Caruso's argument. If it turns out that a causal historical convention *could not* be operative for 'free will', then Nichols' pluralism would fall to eliminativism by default. But this conclusion is not a charitable one, given the strong case that Heller, Nichols himself, and others have made for thinking that a causal historical convention *could* be operative for 'free will', at least in some contexts.[19] For example, even if Caruso is right about the role of our phenomenology of free agency in our initial baptism *and* Nichols is correct that our current phenomenology of free agency is significantly libertarian, this does not entail that the phenomenology is *entirely* libertarian, or that this feature best captures the similarity between the paradigm experiences ostensively targeted by our initial baptism. Even if Nichols is right, surely there are also compatibilist-friendly features of our experience of agency sufficiently similar to demarcate a genuine kind of agency, and one of them might plausibly have been of interest to us at the time of initial baptism. On this point even staunch libertarians are likely to agree, though they will of course deny that this kind of agency is sufficient to ground basic-desert-entailing moral responsibility. Only the most extreme skeptics[20] will deny that these more minimal kinds of agency *exist,* and are the kind of thing that creatures like us might naturally try to talk about.

The upshot here is that on one hand Caruso's argument is subject to the charge of employing an uncharitable interpretation of the conditions that render a causal historical reference-fixing convention operative. Without sneaking in libertarian content we have no reason to think that the initial baptism of 'free will' will fail to pick out a genuine kind in the way that he suggests.

[19] See also Deery (2021b).

[20] Perhaps an epiphenomenalist who believes that all mental events are caused by physical events and no mental events are among the causes of any physical events might wish to deny that even such minimal forms of agency exist. See, for example, Wegner (2002, 2008), and Libet's (1985, 2004) work is sometimes taken to support versions of epiphenomenalism. See also the exclusion principle introduced by Malcom (1968) and further developed by Kim (1989).

4.2 A Better Account of Baptism

However, let us assume for the sake of argument that Caruso is right about initial baptism. Let's say that a causal historical reference-fixing convention is operative for 'free will', that the *only* kind of agency capable of unifying our paradigm experiences of free agency as a genuine kind at the time of initial baptism is libertarian agency, and yet we have come to discover that we never actually instantiate such agency. Even so, we need not conclude that 'free will' fails to refer.

To see why, consider how things might proceed from our initial baptism onward. On this picture the initial baptism of 'free will' would again proceed via some manner of ostensive introspection. We introduce the term 'free will' to talk about the kind of agency we exercise when we experience our first-personal sense of freedom, the only robust similarity between the paradigms of these experiences tracks libertarian features of agency, but we come to discover that there are necessary conditions for this kind of agency that creatures like us do not (or cannot) satisfy. On this picture (keeping in mind that it emerges only after granting a rather long list of assumptions) our initial baptism of 'free will' would fail to secure reference. However, this fact alone does not entail that we fail to refer with our *current usage* as well.

The problem for Caruso here is that he is overlooking one possible – if not prominent[21] – path to preservationism. Preservationists can *grant* that our initial baptism of 'free will' failed to successfully fix the extension of the term, and that this initial baptism (as in the Becher example above) involved a swing and a miss. Even so, preservationists can still argue that this shows only that *we should change the subject*. Unlike the phlogiston case, perhaps there is some very closely related kind of agency that *is* instantiated by creatures like us, can *also* unify a subset of our paradigmatic experiences of free agency as a genuine kind, and that we have clear *interests* in trying to talk about. If we come to find out that the kind of thing we were aiming at with our initial baptism turns out not to be a genuine kind, but that there is a nearby kind sufficiently similar, then it is open to preservationists to argue that a kind of *re-baptism* is in order. Importantly, this sort of move would acknowledge that eliminativists are getting something right – our initial baptism of 'free will' failed to secure reference. But

[21] This is somewhat surprising given that Manuel Vargas (2011, 2013) has explicitly identified and defended this kind of path to preservationism at length. While Vargas refers to this variety of preservationism as *denotational revisionism* it may also be helpful to think of it as a form of *replacementism* (see McCormick (2017)).

they are also getting something wrong – this failure does not entail *current* reference failure and recommend eliminativism, because we may have plausibly *replaced* the empty extension with a second baptism that does successfully fix reference to a genuine kind. Instead of reference failure this would amount to a kind of referential *shift* over time, and one that a causal historical reference-fixing convention is particularly hospitable to.[22]

This variety of preservation can grant that our initial attempt at baptism for 'free will' got things so wrong that it failed to fix reference. But (and perhaps I've sufficiently beaten a dead horse in terms of the baseball analogy at this point) why think that our success at fixing the reference of a term is a one-strike-only-affair? Surely it is not the case that every time we try to pick out a feature of the world with a new term and get things wrong, that term is suddenly off limits in perpetuity. It should come as no great surprise that, for an important feature of human life stretching so far back as free will, we will have gotten things pretty terribly wrong *at least* once. But that in itself does not entail current (or even future) failure. While Caruso seems to assume that successfully fixing reference is a one-strike-and-you're-out affair, this is a particularly implausible way to think about how reference might evolve over time, especially when a causal historical reference-fixing convention is operative.

Taken together, these considerations significantly undermine the force of Caruso's discretionary case for eliminativism. On one hand, claiming that our initial baptism of 'free will' aims at specifically libertarian agency runs the risk of illicitly sneaking conceptual content into a purportedly causal historical reference-fixing picture. On the other hand, holding fixed a more charitable picture of how a genuinely causal historical reference-fixing convention might work for 'free will' while granting a wide swath of Caruso's further claims about initial baptism *still* fails to entail reference failure for our current use of the term 'free will'. Caruso is in need of further argument that this initial baptismal failure

[22] Some readers may find any suggestion that we *should* change the subject initially puzzling, given that charges of subject-changing in philosophical domains like this one are often lodged as objections. But here it is worth noting that not all instances of changing the subject are of a piece. The charge of subject-changing is most troubling for a view, for example, when the shift in question occurs implicitly, or under the guise of the claim that we are all talking about the same thing. But the kind of referential shift suggested here could only succeed by first explicitly acknowledging our previous failure to talk about the same thing, as in the phlogiston case. So long as the shift is acknowledged and carefully argued for, I see no reason to think that it is problematic, nor that it would fail to preserve intertheoretical discourse. Thanks to an anonymous referee for pressing this point.

somehow blocks any subsequent attempts to shift reference and pick out a genuine kind. And I am not optimistic about the prospects for such an argument given the reality of human linguistic flexibility over time.

5. A Discretionary Case for Preservationism

Where does all this leave the disagreement between preservationists and eliminativists about free will? It seems to me, first, that even some variety of pluralism would be a victory for preservationists. To the extent that there are *any* contexts in which the claim, "Free will exists" turns out to be true, preservationists will technically have won the day. However, I think that an even stronger case for preservationism emerges from the various insights about the way that 'free will' refers discussed thus far. It is crucial for eliminativists to make a case for complete and current reference failure, and while the arguments in the previous section show how one explicit attempt to do so fails, preservationists can and should say more. Here I will attempt to do so by continuing to build on the scaffold that Nichols' discretionary view provides. While Caruso's attempt to use this view to motivate eliminativism accepts that both a descriptive and causal historical reference convention can be appropriate in fixing the reference of 'free will' in certain contexts, here I will argue that our all-things-considered interests render one of these possible operative conventions *more appropriate* than the other. And so we should abandon Nichols pluralism and adopt the more appropriate preservationist convention across the board.

5.1 The Problem with Pluralism

I'll begin by motivating a serious problem for pluralism utilizing one of Nichols' own examples, the term 'witch' in a high stakes historical context such as 16th century Salem, Massachusetts (Nichols, 2015: 67). In this context it seems clear that a descriptive reference-fixing convention (for example, the extension of 'witches' is picked out descriptively by something like, "has a pact with Satan") could be operative. The sort of practical interests that might render a descriptive convention operative in this context might involve, say, concerns about protecting the moral community from agents of the devil while also making sure not to burn innocent people who have no such pacts alive. Given these interests we ought to err on the side of caution when it comes to making existence claims about witches that might easily get people killed. In this context a descriptive

convention could be operative, result in reference failure, and recommend eliminativism about witches. No one satisfies the relevant description (having a pact with Satan), and so we ought to conclude that witches do not exist, stop claiming that they do, and – most importantly – stop burning people at the stake based on misattributions of witch-hood.

Consider, though, what happens if we extend Nichols' pluralism about 'free will' to the term 'witch', holding fixed the same historical context. Like philosophers engaged in debates about free will it is unlikely that everyone's motivating concerns about witches will be the same, even in the same historical context or moral community. 16th century scholars, perhaps, might be interested in long-standing pagan traditions, especially those involving knowledge of natural medicinal resources. Their dominant motivating concern in witches might be to be able to identify and consult them in order to help the community by expanding our knowledge of the possible treatments for disease. And if 'witch' is an ambiguous kind term then this motivating concern could render a liberal causal historical reference-fixing convention operative for a scholar in this context. For such scholars the term 'witch' ostensively aims at a certain kind of person (the kind, say, who has a knack for healing based on their knowledge of pagan traditions), there is sufficient similarity between paradigms, and so when such a scholar says, "Witches exist" he might plausibly speak truly.

Such is the case if we embrace Nichols' pluralism for ambiguous kind terms, and 'witch' turns out to be one of them. But now we have a potentially serious problem. In the historical context we've been considering the descriptive convention results in reference failure and the causal historical convention allows for successful reference. When, for example, the friend of an accused witch (let's call her Anne) says, "Witches don't exist!" one of the dominant motivating concerns at play is to make sure that innocent people – especially Anne! – are not burned at the stake. This motivating concern plausibly renders the descriptive convention operative, and what Anne's friend says is true. Witches don't exist, and so Anne is not a witch. But, let's say that Anne is also the kind of person that a scholar sketched above is interested in. Perhaps she is the village healer, and has a wealth of knowledge passed down to her by generations of women who have studied pagan traditions regarding natural medicinal resources. If 'witch' is an ambiguous kind term and we accept Nichols' pluralism, then when the scholar says, "Witches do exist, and Anne is one of them," a causal historical convention is plausibly operative and what he says is *also* true. So far so good for pluralism, but not so much for Anne. Despite the scholar's

genuinely valuable motivating concern about witches (to benefit the community via improved medicinal knowledge) his utterance could easily get her killed.

From our contemporary vantage point, what Anne's example shows is that while both kinds of motivating concern about witches are worth caring about, they are not *equally* so. In fact, I submit that *we* s*hould care more*, all-things-considered, about preventing innocent people from being burned alive than we should about benefiting our community with expanded medicinal knowledge. This is not to say that the latter concern is not important, just that it is obviously and uncontroversially not *as* important as making sure we don't burn innocent people alive.[23] And for those readers who disagree, consider how one might make a case for this position *to Anne*. Here of course one might appeal to explicitly consequentialist considerations and argue that acquiring the relevant medicinal knowledge would obviously have more overall utility than saving Anne's life (and perhaps even the lives of the comparative handful of other women identified as witches), but this of course will be cold comfort to Anne and her loved ones.

Furthermore, eliminativists are often unwilling to embrace this kind of strictly consequentialist motivating concern. For example, Pereboom (2020) and Caruso (2021) each attempt to argue that the justification for eliminativist quarantine models of criminal punishment can be supported by deontological principles (especially the right to self-defense) alone, largely in order to avoid worries about preemptive incapacitation and the use objection that inevitably crop up as a result of reliance on consequentialist principles to do this work. Perhaps, then, a further merit of identifying the important role that motivating concerns play in this debate is that it suggests further that eliminativists may want to reconsider this current disavowal of consequentialist motivating principles. Embracing and defending them might in fact be one of the most plausible

[23] I am not presuming that embracing eliminativism is the only way to end the practice of burning innocent people alive as witches in this context. Perhaps, for example, 16th century villagers could maintain that witches exist but jettison the widespread belief that they have a pact with the devil, or even adjust their beliefs about the moral implications of such a pact. While conceivable, though, such scenarios strike me as highly unlikely holding fixed the historical context that we are considering. After all, this is precisely the sort of thing that many villagers concerned about the practice of burning their neighbors alive surely *tried* to initiate, alas without much success. Were I to find myself in the shoes of the accused I would certainly prefer my defenders to wax eliminativist, rather than *grant* that I am a witch while proceeding to argue that being one is not so bad as it seems. Thanks to an anonymous referee for raising this question.

paths to arguing that we *should* care more about the concerns motivating eliminativists than those motivating preservationists. This will of course be a bullet to bite, but surely those who defend high stakes varieties of eliminativism expect to bite some hefty bullets along the way.

Returning to the example at hand, if I am correct that we should care more about avoiding burning innocent people alive than increasing our medicinal knowledge then it looks as though we should reject pluralism about the term 'witch' in the historical context we've been considering, even if 'witch' is a genuinely ambiguous kind term. In this context we ought to adopt the descriptive convention rendered operative by the weightier motivating concern. Even if a causal historical convention is available for some speakers, they *should not* use the term 'witch' in accordance with this convention. Anne's example helps to show that in circumstances like this the descriptive convention is obviously *more appropriate*, and the one that should govern our assessment of truth of claims about witches in this high stakes context.

I think that a similar line of argument can and should be made for 'free will' in the context of the contemporary debate between eliminativists and preservationists. Tokens of the term 'free will' are uttered in circumstances bound up in similarly weighty motivating concerns as the term 'witch' above, especially when we are trying to decide whether someone is blameworthy in both interpersonal and criminal contexts. I will now turn to the task of making the relevant motivating concerns explicit for both eliminativists and preservationists about free will, and argue that we should, all-things-considered, care more about the concerns that motivate preservationists.

5.2 Preservationism and Prioritizing Victims

One feature of the disagreement between preservationists and eliminativists about free will that I have long found fascinating is that both sides appear to be motivated by powerful, yet largely distinct motivating moral concerns. First, the dominant motivating concern cited by eliminativists is often a concern about *widespread undeserved harm*. We can all agree that harm is bad, that undeserved harm is especially so, and that our responsibility-related attitudes and practices often cause a great deal of harm. This is true in regard to both our interpersonal practices of blaming and holding one another responsible – especially when it comes to angry, reactive blame – and in regard to our legal practices and retributive systems of criminal punishment. Eliminativists are worried that the justification for all of these harms depends crucially on the claim that they are deserved

in the basic sense by those subjected to them, but, if 'free will' fails to refer then such claims are always false. Therefore maintaining our responsibility-related attitudes and practices generates widespread undeserved harm.

When we focus on eliminativist concerns about widespread undeserved harm the contemporary free will debate looks similar to Anne's context above in that the moral stakes are high. Our current system of retributive criminal punishment, for example, does not allow us to burn criminals at the stake but it does allow for other means of state sanctioned killing and a wide array of inhumane detention conditions. And our interpersonal blaming practices may not be as severe as the kinds of social consequences Anne might have experienced if accused of being a witch, but they do often justify a wide array of deeply unpleasant experiences such as being the target of resentment, guilt, and indignation as well as withdrawal from, damage to, and sometimes even the destruction of some of our most deeply valued interpersonal relationships.

It is no great surprise then that, like Anne's friends and loved ones, eliminativists think we ought to err on the side of caution in regard to what counts as free will.[24] Eliminativists are trying to talk about the kind of thing that could possibly justify all of this widespread harm, and so if 'free will' is in fact an ambiguous kind term a descriptive reference-fixing convention is plausibly operative for them. We need to appeal to some conceptual content – something like ultimate sourcehood, perhaps – in order to make sense of the kind of agency that could at least render all of this harm deserved. And, if Nichols and others are correct and we have no good reason to think that we ever actually instantiate this kind of agency, then when eliminativists say that "Free will doesn't exist" they speak truly.

But this is not the full story when it comes to 'free will'. We can – and in fact I think we should – grant that eliminativists' motivating concerns are powerful, morally speaking. *Of course* we ought to care, and deeply so, about avoiding widespread undeserved harm. However, there is also an important difference between the context of the contemporary free will debate and Anne's context. In Anne's case, *failing* to attribute witch-hood to Anne (or anyone else) has minimal *negative* consequences. While the scholar we considered above has noble motives in wanting to benefit his community by acquiring new medicinal knowledge, finding a witch whose expertise he can consult is not the only way to achieve this goal. The scholar might consult alternative sources. And even if he

[24] And thus in turn basic-desert-entailing moral responsibility and blameworthiness.

is dead set on learning more about the kind of knowledge folks like Anne possess, he can easily consult Anne's expertise without attributing to her kind-membership that might also get her killed. Perhaps, for example, in light of the danger that calling people witches entails the scholar realizes that it would be better to introduce a new term altogether (perhaps 'healer' or 'doctor' would be apt here) to identify people like Anne for his purposes. But there is no need – or even a *good reason* – for him to call her a witch, even if 'witch' is an ambiguous kind term and doing so would generate a true utterance.

Here I submit that, unlike witch-hood, *failures* to attribute free will and in turn hold one another responsible can have overwhelming negative consequences. When the scholar refrains from calling Anne a witch *there are no victims*. But when it comes to a subset of the actions we are interested in when we talk about free will – those that involve some kind of harm – there is *always* a victim. Whether or not we attribute free will and hold those who have harmed them responsible is not some ho hum affair for victims, and *protecting and defending* them is one of the primary motivating concerns for preservationists when it comes to their interests in theorizing about free will. This is not to say that preservationists do not care about undeserved harm, only that they care *more* about the way that failures to attribute free will, blame, and hold one another responsible can negatively impact the victims of harms that have already occurred.[25]

While considerations of space prohibit a lengthy discussion of the role that our attributions of free will and responsibility-related attitudes and practices play in protecting and defending victims, here I will assume that those who argue that wrongful treatment often carries with it an implicit claim about the value of the subject of that treatment are correct. Failures to challenge this claim via our responsibility-related attitudes and practices can work to *confirm* this implicit claim, and can even disvalue the victim further.[26] Returning to reference, if 'free will' is an ambiguous kind term then motivating concerns about the

[25] Nor is this to say that eliminativists do not care *at all* about protecting and defending victims. Of course they do, and this is precisely why the most prominent among them offer some kind of elimination-friendly model of criminal punishment to replace our current retributive, basic-desert entailing system (see, for example Caruso (2021) and Pereboom's (2001, 2014) quarantine models, and Waller's (2019) support of restorative models).

[26] For further discussion of the way that wrongful treatment can carry with it various kinds of meaning in regard to the value of the victim see Hieronymi (2004), Franklin (2013), and Smith (2013).

need to protect and defend victims could plausibly render a causal historical reference-fixing convention operative for preservationists. What they are interested in when talking about free will is the kind of action that carries with it implicit claims about the value of other agents, especially claims that *disvalue* other agents. Because a wide swath of our actions are likely to share relevant similarities with paradigms of this kind, and there are no obviously exotic metaphysical conditions on our actions carrying this kind of meaning, when preservationists say "Free will exists" what they say is also plausibly true.

Or so, again, this is where pluralism would land us. But as we saw with Anne's case above there is a serious problem for pluralism about ambiguous kind terms when the motivating concerns that render one reference-fixing convention operative are *obviously weightier* that those that render another operative. Is this the case for 'free will' in the context of the contemporary debate between preservationists and eliminativists? I think that it is, but in order to see why a final toy case will be helpful. Consider the following:

> **Vlad:** Vlad is a vicious autocrat hell bent on invading and acquiring the territory of one of his neighboring sovereign nations. Armed with a stockpile of nuclear weapons, Vlad is confident that other nations will stay out of the fray, and that he will succeed. But Vlad is wrong. The citizens of the nation he invades put up a heroic fight, and the invasion ultimately fails. Furthermore, it comes to light that Vlad's gruesome tactics knew no bounds – he intentionally targeted civilians (children's hospitals, maternity wards, and schools all being commonplace) with air strikes and chemical weapons resulting in the deaths of thousands of innocent children and civilians.

Luckily, Vlad has recently been captured by enemy forces and is being put on trial internationally for war crimes. Many victims of the conflict will be given an opportunity to testify, and one is a woman named Yulia. Yulia lost her only child in an airstrike, and was also viscously physically and sexually assaulted by a contingent of Vlad's troops, all of whom received explicit orders (stemming from Vlad himself) encouraging them to engage in such behavior. At the end of her tearful testimony Yulia implores the tribunal to hold Vlad responsible. "He is a monster," she says, "he freely ordered this kind of behavior, and he is to blame for all of the misery that I have endured."

Yulia is the final victim to testify, and after a short recess Vlad's lawyers begin mounting a defense in their opening statement. Their central argument involves appeal to global skepticism about free will and moral responsibility. They argue that Vlad was not responsible for his genes and his upbringing within

a similarly brutal regime. While his actions were, of course, abhorrent, they were not "up to him" in a deep sense. In the kill-or-be-killed circumstances he found himself in we could not expect more from Vlad, morally speaking. "Really, Vlad is a survivor himself," say his lawyers, "and while the harms experienced by these many witnesses are regrettable, ultimately Vlad is not free and responsible, at least not in the basic-desert-entailing sense necessary to justify the punishment that this tribunal wishes to inflict on him."[27]

First, what conclusions would discretionary pluralism about 'free will' generate in this case? Regarding Yulia's utterance pluralism would likely entail that what Yulia says is true. As discussed above one set of motivating concerns in talking about free will and responsibility is to protect and defend victims, and in this particular case the victims of Vlad's violent invasion. This motivating concern would render a causal historical reference-fixing convention operative for Yulia, and the paradigm actions aimed at would be the kind that carry with them negative implicit claims about the value of victims of harm. Insofar as there are such actions (it seems clear that there are) and the subset of Vlad's actions that have harmed Yulia are of this kind (it seems clear that they do), then what Yulia says is true.

Turning to the lawyer's utterance, another motivating concern at play in this context is the eliminativist concern about undeserved harm. This concern would again plausibly render a descriptive reference-fixing convention operative, and the relevant description would be the kind of libertarian agency needed to secure basic-desert-entailing responsibility alluded to in various comments from Vlad's lawyer above. But, if skeptics are right then we never actually instantiate this kind of agency, 'free will' fails to refer, and the lawyer's utterance that "Vlad is not free and responsible" is true along with any further claims that this entails about the lack of justification for basic-desert-dependent punishment. And so, according to pluralism, what Vlad's lawyer says is also true.

If we embrace pluralism then this is the end of the story about how to assess Vlad's case. What Yulia says and what the lawyer says are *both* true claims.

[27] Is this case too extreme or rare to tell us much about our actual responsibility-related practices and the contexts in which they occur? I think not. Setting aside the fact that the case is already based somewhat on current events, I think that our feelings about Yulia here should generalize to any other victim of sexual assault or the parent of a murdered child. While I wish such cases were rare, unfortunately they are not, and part of the case I hope to be making here is that our theorizing about free will and moral responsibility should not proceed as if they are. Thanks to an anonymous referee for pressing this point.

I think, first, that this result is itself sufficient to motivate our rejection of pluralism in high stakes contexts like this one, and those involving attributions of free will and moral responsibility more generally. What, after all, is the *tribunal* to do with this result? They must render a verdict on whether or not Vlad is responsible for committing war crimes and should be punished accordingly. To allow that both Yulia and Vlad's lawyer speak truly is to commit to a verdict that will either involve punishing Vlad (likely severely) undeservedly, or to allow a legitimate moral monster to go free. Insofar as such verdicts and broader legal practices can be sustained at all "He is *both* innocent and guilty," will never be a viable option.

So then what is the tribunal to do? I think that the answer here is obvious – the tribunal should side with Yulia. Even if concerns about avoiding undeserved harm are very important, morally speaking, and Vlad's lawyers are correct that we never exercise the kind of libertarian agency that could render such harm deserved in the basic sense, it just seems painfully clear that we ought to *care more* about protecting and defending Yulia and the thousands of other victims like her than our (albiet also well founded) concerns about potentially harming wrongdoers like Vlad when they may not deserve such treatment in the basic sense.

At this point it will come as no surprise to the reader that I see the tribunal's role as parallel to our own in the broader context of contemporary debates about free will and moral responsibility. When adjudicating between preservationists and eliminativists, we ought to pay closer attention to the moral concerns that *motivate* each of these positions, more clearly *articulate* the role that these concerns play in determining the reference of 'free will', *acknowledge* the moral significance of both kinds of concerns, and ultimately take a stand on which moral concern *we should care about more*. And I submit that we should, all-things-considered, care more about protecting and defending victims than avoiding the potentially undeserved harms for wrongdoers embedded in our responsibility-related attitudes and practices.

Why? While this weighting of moral concerns admittedly feels obvious to me, here of course I anticipate that at least some readers – especially those with eliminativist leanings – will disagree. For such readers my concluding remarks will likely be unsatisfying, but I would recommend ruminating more seriously on Yulia's case and countless others involving victims in our actual lives. Eliminativists have much to say about the benefits of letting those who do not deserve blame – even folks like Vlad – off the hook. But they rarely discuss how we ought to break the news that we intend to do so to their victims. For those

eliminativists still reading, what should we say to Yulia if Vlad goes unpunished? I myself am at a genuine loss when considering this question. I can see no good way to attempt to justify this result to Yulia in a way that does not feel like a further assault on her value. Perhaps I simply lack sufficient imagination, but insofar as eliminativists wish to maintain that we should care more about concerns about undeserved harm than protecting and defending victims I leave the question of how to explain this position to actual victims as an open one.

6. Conclusion

While much more needs to be said about how we ought to weigh the motivating concerns of eliminativists and preservationists, here I hope to have moved the methodological dial forward a bit by articulating the way that issues concerning reference impact this debate. What we discover when we look more carefully at the role of reference-fixing conventions in the contemporary debate between eliminativists and preservationists in particular is that the motivating moral concerns of these two camps play an important yet thus far largely overlooked role. And with a better understanding of the relevant mechanisms in hand we are better positioned to both diagnose some of the apparently intractable character of this disagreement (both camps are motivated by moral concerns worth caring about) and engage directly with the question of which concerns we ultimately ought to care about more.

REFERENCES

Andreasen, R. (2000). Race: Biological Reality or Social Construct? *Philosophy of Science*, 67, S653S666.

Appiah, K.A. (1995). The Uncompleted Argument: Du Bois and the Illusion of Race. In L.A. Bell and D. Blumefeld, eds., *Overcoming Racism and Sexism*. Lanham, MD: Rowman & Littlefieled, pp. 59-78.

Blackburn, S. (1985). Errors and the Phenomenology of Value. In T. Carson and P. Moser, eds., *Morality and the Good Life*. New York: Oxford University Press, pp. 324-337.

Campbell, J. (1957). *Of Selfhood and Godhood*. London: Allen & Unwin.

Caruso, G. (2015). Free Will Eliminativism: Reference, Error, and Phenomenology. *Philosophical Studies*, 172, 2823-2833.

Caruso, G. (2021). *Rejecting Retributivism: Free Will, Punishment, and Criminal Justice*. Cambridge: Cambridge University Press.

Churchland, P.M. (1981). Eliminative Materialism and Propositional Attitudes. *Journal ofPhilosophy*, 78, 67-90.

Churchland, P.S. (1986). *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. Cambridge, MA: MIT Press.

Daw, R., and Torin Alter. (2001). Free Acts and Robot Cats. *Philosophical Studies*, 102, 345-357.

Deery, O., Matt Bedke, and Shaun Nichols. (2013). Phenomenal Abilities: Incompatibilism and the Experience of Agency. In D. Shoemaker, ed., *Oxford Studies in Agency and Responsibility*. New York: Oxford University press, pp. 126-150.

Deery, O. (2021a). Free Actions as a Natural Kind. *Synthese*, 198(1), 823-843.

Deery, O. (2021b). *Naturally Free Actions*. Oxford: Oxford University Press.

Feyeraband, P. (1962). Explanation, Reduction and Empiricism. In H. Feigel and G. Maxwell, eds., *Scientific Explanation, Space, and Time*, vol. 3 of Minnesota Studies in the Philosophy of Science. Minneapolis, MN: University of Minnesota Press, pp. 28-97.

Fodor, J. (1998). *Concepts: Where Cognitive Science Went Wrong*. New York, NY: Oxford University Press.

Franklin, C. (2013). Valuing Blame. In D.J. Coates and N. Tognazzini, eds., *Blame: Its Nature and Norms*. New York, NY: Oxford University Press, pp. 207-223.

Griffiths, P. (1997). *What Emotions Really Are*. Chicago, IL: University of Chicago Press.

Heller, M. (1996). The Mad Scientist Meets the Robot Cats: Compatibilism, Kinds, and Counterexamples. *Philosophy and Phenomenological Research*, 56(2), 333-337.

Hieronymi, P. (2004). The Force and Fairness of Blame. *Philosophical Perspectives*, 18, 115-148.

Hurley, S. (2000). Is Responsibility Essentially Impossible? *Philosophical Studies*, 99(2), 229-268.

Kane, R. (2017). Free Will Bound and Unbound: Reflections on Shaun Nichols' *Bound*. *Philosophical Studies*, 174(10), 2479-2488.

Kim, J. (1989). Mechanism, Purpose, and Explanatory Exclusion. *Nous-Supplement: PhilosophicalPerspectives*, 3, 77-108.

Kripke, S. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Laudan, L. (1984). *Science and Values*. Berkeley: University of California Press.

Lewis, D. (1972). Psychophysical and Theoretical Identifications. *Australasian Journal ofPhilosophy*, 50, 249-258.

Libet, B. (1985). Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action. *Behavioral and Brain Sciences*, 8, 529-566.

Libet, B. (1992). The Neural Time-Factor in Perception, Volition and Free Will. *Revue de Métaphysique et de Morale*, 2, 255-272.

Machery, E. (2009). *Doing without Concepts*. Oxford: Oxford University Press.

Mackie, J. (1977). *Ethics: Inventing Right and Wrong*: New York, NY: Penguin.

Malcom, N. (1968). The Conceivability of Mechanism. *Philosophical Review*, 77, 45-72.

Mallon, R. (2006). Race: Normative, Not Metaphysical or Semantic. *Ethics*, 116(3), 525-551.

McCormick, K. (2017). Revisionism. In N. Levy, M. Griffith, and K. Timpe, eds., *The Routledge Companion to Free Will*. New York, NY: Routledge, pp. 109-120.

McCormick, K. (2019). Meeting the Eliminativist Burden. *Social Philosophy and Policy*, 36(1), 132 153.

McCormick, K. (2022). *The Problem of Blame: Making Sense of Moral Anger*. Cambridge: Cambridge University Press.

McKenna, M. (2009). Compatibilism and Desert: Critical Comments on *Four Views on Free Will*. *Philosophical Studies*, 144, 3-13.

Nichols, S. (2015). *Bound*. Oxford: Oxford University Press.

Nichols, S. (2017). Replies to Kane, McCormick, and Vargas. *Philosophical Studies*, 174(10), 25112523.

O'Connor, T. (1995). Agent Causation. In T. O'Connor, ed., *Agents, Causes and Events: Essays on Free Will and Indeterminism*. Oxford: Oxford University Press, pp. 173-200.

Pereboom, D. (2001). *Living without Free Will*. Cambridge: Cambridge University Press.

Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.

Pereboom, D. (2020). Free Will Skepticism, General Deterrence, and the "Use" Objection. In E. Shaw, D. Pereboom, and G.D. Caruso, eds., *Free Will Skepticism in Law and Society: Challenging Retributive Justice*. Cambridge: Cambridge University Press, pp. 91-115.

Putnam, H. (1962). It Ain't Necessarily So. *Journal of Philosophy*, 59, 658-671.

Putnam, H. (1975). The Meaning of "Meaning". *Minnesota Studies in Philosophy of Science*, 7, 131193.

Smith, A. (2013). Moral Blame and Moral Protest. In D.J. Coates and N. Tognazzini, eds., *Blame: Its Nature and Norms*. New York, NY: Oxford University Press, pp. 27-48.

Stitch, S. (1983). *From Folk Psychology to Cognitive Science*. Cambridge, MA: Cambridge University Press.

Strawson, G. (1994). The Impossibility of Moral Responsibility. *Philosophical Studies*, 75, 5-24.

Taylor, R. (1992). *Metaphysics*, 4th ed. Englewood Cliffs, NJ: Prentice-Hall.

Unger, P. (1979). There are No Ordinary Things. *Synthese*, 41, 117-154.

Van Inwagen, P. (1990). *Material Beings*. Ithaca, NY: Cornell University Press.

Vargas, M. (2011). Revisionist Accounts of Free Will: Origins, Varieties, and Challenges. In R. Kane, ed., *The Oxford Handbook of Free Will*, 2nd ed. Oxford: Oxford University Press, pp. 457-474.

Vargas, M. (2013). *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.

Vargas, M. (2017). Contested Terms and Philosophical Debates. *Philosophical Studies*, 174(10), 2499-2510.

Waller, B. (2019). Beyond the Retributive System. In E. Shaw, D. Pereboom, and G. Caruso, eds., *Free Will Skepticism in Law and Society: Challenging Retributive Justice*. Cambridge: Cambridge University Press.

Wegner, D. (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Wegner, D. (2008). Self is Magic. In J. Baer, J.C. Kaufman, and R.F. Baumeister, eds., *Are We Free? Psychology and Free Will*. New York, NY: Oxford University Press.