Free actions as a natural kind

Oisín Deery

1. Introduction

Does free will exist? Understanding free will as the ability to act freely, and free actions as exercises of this ability, I maintain that the default answer to this question is "yes," by maintaining that FREE ACTION is a natural-kind concept and free actions are a natural kind.[1] The resulting position builds on the view that *agents* are a natural kind (Sims 2018) and yields an attractive alternative to recent revisionist treatments of the concept FREE ACTION (Heller 1996; Vargas 2013; Nichols 2015).

2. Free will and reference

In recent years, a lively debate has emerged that addresses the question whether free actions exist, by asking whether the concept FREE ACTION refers (Heller 1996; Hurley 2000; Vargas 2013; Nichols 2015; Caruso 2015; Deery 2015a; McCormick 2016, forthcoming; cf. Sims 2018). On this approach, if FREE ACTION refers to any of our behaviors, then free actions exist; otherwise not.[2] The ensuing debate focuses on whether we can preserve the concept FREE ACTION rather than

[1] I will use capitals for concepts (e.g., FREE ACTION) and single quotation marks for terms (e.g., 'free action'), where relevant.

[2] The move from considerations about reference to conclusions about existence has been challenged (e.g., Mallon et al. 2009).

eliminate it, on the assumption that it is associated with errors.[3] The most commonly suggested error is that the concept falsely presupposes indeterminism as being usefully implicated in free actions (Vargas 2013, pp. 21–72; Nichols 2015, pp. 54–5; cf. May 2014; Deery et al. 2015). In what follows, we will not assume this particular error but will instead consider how theories might handle errors more generally.[4]

For descriptivists, concepts are treated as analogous to theoretical terms, which refer (if they do) to whatever satisfies a critical set of claims associated with them (Lewis 1972, p. 213). Accordingly, FREE ACTION refers just in case some class of behaviors satisfies the major presuppositions associated with the concept. Adopting a *conservative* descriptivist theory (i.e., one that is relatively intolerant of errors), one might hold that FREE ACTION fails to refer if a major presupposition associated with it is erroneous. As a result, one might be an eliminativist about the concept and deny that free actions exist (Caruso 2015; cf. Strawson 1986; Levy 2011; Pereboom 2014; Nichols 2015).

By contrast, Manuel Vargas thinks that reference might be secured, and the concept preserved, despite various errors, by abandoning the erroneous presuppositions and revising the concept (e.g., 2011, p. 460). Vargas calls this position *revisionism*. In sketching the position, Vargas seemingly adopts a *liberal* descriptivist approach (i.e., one that is tolerant of at least some errors), and he provides a specific revisionist proposal, which is that acting freely is an exercise of

[3] I follow Nichols (forthcoming) in understanding concepts as mental representations, while avoiding commitment to the idea that they contain descriptions. Instead, like Nichols, I adopt "the deliberately vague characterization on which the descriptive information is 'associated' with the concept" (Nichols forthcoming).

[4] Other presuppositions apart from indeterminism might be in error—for example, the presupposition that free actions require conscious or rational control (Levy 2016; Vargas 2017).

the capacity to respond to moral reasons, even if that is not what commonsense presupposes about acting freely.[5]

A very different option is to adopt a liberal yet *non-descriptivist* approach. The canonical approach of this sort is the Putnam-Kripke view, which has two aspects—one semantic, the other metaphysical. Both aspects will be important later, so I will sketch them briefly here.

The semantic aspect concerns how concepts acquire meaning and how reference works for certain concepts, including natural-kind concepts. According to semantic externalism (e.g., Putnam 1975), a concept's referent is the main element of its meaning. Thus, the semantic content of WATER, for example, is specified not in terms of presuppositions associated with the concept, but instead primarily by what water actually is.[6] Additionally, semantic externalists typically adopt the causal-historical theory of reference (e.g., Putnam 1975; Kripke 1980), the aim of which is to explain how a concept comes to be associated with the referent from which it gets its meaning. For kind concepts like WATER, reference is fixed by a baptism, or a dubbing, for instance by perceptual contact with a sample of the kind. The concept then refers to whatever else has the same relevant feature as the dubbed sample, and we continue to use the concept as long as our doing so stretches back in causal

[5] I concede that Vargas might adopt a non-descriptivist natural-kind view (see Section 8.1); however, his specific proposal about free actions goes against the spirit of such views, which typically permit preserving the concept even without specifying any revision, as I explain below.

[6] These claims are typically made in the first instance about linguistic terms. I follow both descriptivists (e.g., Jackson 1998, p. 33) and non-descriptivists (e.g., Laurence and Margolis 2003, p. 257) in maintaining that semantic claims about terms also apply to concepts.

chains to the dubbing. Call the combination of semantic externalism and the causal-historical theory of reference *causal-historical externalism*.

The metaphysical aspect of the Putnam-Kripke view concerns what it is in virtue of which a collection of objects is a natural kind. According to *essentialism*, which tends to go hand-in-hand with causal-historical externalism, the feature that defines a kind is supposed to be both essential and unique to its members.[7] Essentialism, as a metaphysical view, relates to the semantic issue since it says that reference for a kind concept is determined by the essential nature, if any, that paradigm cases of the putative kind have in common. Establishing that essential nature becomes an empirical matter. Following others (e.g., Daw and Alter 2001, p. 347), I call the combination of causal-historical externalism and essentialism *the Putnam-Kripke view*.

Shaun Nichols labels non-descriptivist views of this general sort "placeholder preservationism" (forthcoming), since they permit preserving the target concept and maintaining that it refers to a kind despite the fact that no revision need be specified, by contrast with Vargas's view, which specifies a revision.

The Putnam-Kripke view is not the only (or best) non-descriptivist view available. One could, for instance, endorse causal-historical externalism *without* also endorsing essentialism about kinds. Alternatively, one could reject both these views and endorse (as I will) an entirely different non-descriptivist theory of reference, with a non-essentialist metaphysics. On this view, natural kinds are homeostatic property

[7] Salmon (1982, pp. 161–175) denies that essentialism is entailed by the causal-historical view. Even so, those who adopt causal-historical externalism usually also have essentialist intuitions, and arguments for causal-historical externalism can be framed in non-semantic terms that might be used to motivate essentialism more directly.

clusters, and kind concepts refer to these clusters. I maintain that this highly influential non-descriptivist account provides for a more attractive form of placeholder preservationism about free actions than the Putnam-Kripke view.

3. Homeostatic property clusters

According to Richard Boyd's (1988, 1999) homeostatic-property-cluster (HPC) theory of natural-kinds, for any collection of objects to count as a natural kind, the objects composing the collection must share a homeostatic cluster of properties that contingently co-occur. The cluster should be homeostatic in the sense that the presence of some of the properties favors the presence of others, or an underlying mechanism maintains all of them together. As a result, their co-occurrence, although contingent, is not accidental; it is due to mechanisms that causally maintain it.

Additionally, the cluster must be causally relevant, in that it should identify causally sustained generalizations that achieve the inductive-explanatory (or other) goals of an inferential domain *because* the occurrence of (some of) the properties, together with the relevant homeostatic mechanisms, reliably results in certain effects, or facilitates our achieving certain results. In this way, the cluster "accommodates" (Boyd 1999, pp. 141–85) the demands of the relevant domain by serving our goals in tracking it. Kinds are in this way projectable, since they permit legitimate inductive inferences; yet they are not merely nominal, since they are tied to causal structure (Wilson et al. 2007).

A concept, C, that applies to a homeostatic cluster is defined by all (or part) of the cluster, together with its mechanisms of homeostasis (Boyd 1999, pp. 143–4; cf. Brigandt 2011). The causal importance of the cluster is such that the kind picked

out by C is legitimately a natural kind. Thus, ORANGUTAN refers to a natural kind—a biological species—since orangutans share a cluster of properties that is causally maintained in the right sort of way. The mechanisms maintaining homeostasis need not, however, be *intrinsic* to members of a kind. For instance, the evolutionary niche occupied by orangutans has presumably resulted in selection pressure for specific traits; but this mechanism is *extrinsic* to members of the kind. Nevertheless, it enables us to identify fruitful causal-explanatory generalizations.[8]

Boyd's aim in developing the HPC view was partly to provide an account of how species are natural kinds. On the Putnam-Kripke view, the feature that defines a kind must be essential and unique to its members. However, it is notoriously difficult to find biological traits that occur in all and only the members of a species; mutation can result in the disappearance of a trait in a species' future members, while the disappearance of a trait in even one member is enough to show that the trait is not essential to that species. Moreover, organisms in different species might have common traits, and so these traits are not unique to a particular species. These problems disappear on the HPC view, since no single feature in a homeostatic cluster is itself necessary to a kind; any feature might disappear, yet as long as the cluster remains homeostatic it will continue to define the kind. Likewise, any feature in a homeostatic cluster might be possessed by members of other kinds, with the result that it need not be unique to members of a particular kind.

Boyd's account is now the received view of natural kinds in the philosophy of science (Ereshefsky and Reydon 2015). It has also been applied in metaphysics

[8] The mechanisms underlying homeostasis might also be historical. Thus, a biological species (like orangutans) is partly defined in terms of a cluster of properties that is homeostatic due to common evolutionary descent (Boyd 1999, p. 144, pp. 154–6).

(Khalidi 2018; cf. Wilson et al. 2007), in relation to moral issues (Pober 2013; Kumar 2015; Deery 2015a; Sims 2018), in epistemology (Kumar 2014), and perhaps most notably in relation to personal identity (Schechtman 2014). On the HPC view, FREE ACTION might refer, and so free actions might exist, even if our presuppositions about the features that underpin free actions are mistaken—perhaps wildly mistaken. The concept refers, and free actions exist, just in case:

(a) the behaviors we naturally categorize as paradigm cases of the (putative) kind have a homeostatic cluster of features in common, and

(b) this cluster serves our goals in tracking it.

As I will explain further in the next section, we naturally categorize certain behaviors in a certain way—as *those* behaviors; they seem to have something in common, even if we do not know what it is. Functionally, they seem to be instances of the most sophisticated goal-directed behavior typically exhibited by human agents, often or even usually in the context of assigning moral responsibility. These behaviors are the (putative) kind's paradigm cases. What supports agreement about these cases and permits us to generalize from them to the kind is (*per hypothesi*) a cluster of features and the usefulness of our tracking them. The nature of the behaviors must be established empirically, as indicated by (a). Our categorizing them as we do, and thereby our acquiring and retaining the concept, is ultimately explained by something akin to what Ruth Millikan calls a concept's "stabilizing function" (2010, pp. 43–81), which is its survival value or usefulness for us, as indicated by (b).[9]

[9] I will return to Millikan's idea of a stabilizing function in Section 7, where I say more about why the HPC view cannot appeal to a criterial or descriptive way of identifying the paradigm cases, or generalizing from them to the kind. Moreover, HPC theorists sometimes add another condition to clauses (a) and (b), such that where the relevant features in (a) are

For ease of reading, and following others (e.g., van Inwagen 1983, pp. 107–8; Heller 1996, pp. 335–6; Daw and Alter 2001, p. 349), I will sometimes rely on the shorthand phrase, "paradigm cases of free action," and I will say things like, "FREE ACTION refers just in case *paradigm cases of free action* have certain features," and so on.[10] In doing so, I am not being circular, once this phrase is understood as outlined above—i.e., as shorthand (roughly) for instances of the most sophisticated goal-directed behavior typically exhibited by humans, often in the context of assigning moral responsibility.[11]

4. FREE ACTION and free actions

Broadly evolutionary considerations speak in favor of the verdict that FREE ACTION is an HPC-kind concept, and free actions are an HPC kind, even without conclusively establishing it.

First, it is overwhelmingly plausible that paradigm cases of free action (understood in the way described above) have in common various features (whatever they are—we do not need to specify what they are in advance) that enable the individuals who possess them to behave in ways that best achieve their goals, across

unspecified, the kind is identified partly by the criteria we *actually* use in identifying kinds for use in explanada, i.e., the factor that makes the kind's members explanatorily interesting, such as the capacity to fulfill a particular functional role (Reydon 2009, pp. 733–744). This extra condition works as a constraint on identifying the (putative) kind in the first place, i.e., on identifying its *paradigm cases*. For me, the relevant functional role is, as I note, that the behaviors in question seem to be instances of the most sophisticated goal-directed agency typically exercised by humans, often in the context of assigning moral responsibility.

[10] Likewise, many who rely on the HPC account for other purposes use a similar shorthand, e.g., Kumar defines reference for 'knowledge' by appeal to "paradigm cases of knowledge" (2014, p. 447, cf. pp. 442–443, 446).

[11] I thank a referee for encouraging me to clarify this point.

variable conditions. Plausibly, these features consist partly in features that enable agents to behave *intelligently* in various domains. From an evolutionary perspective (Sterelny 2003), intelligence broadly conceived is the ability to avoid obstacles or attain ends in a specified domain, and to do so reliably and in differing circumstances. While the domain in question might be abstract and minimally agentive, like mathematics, the *evolution* of intelligence is linked fundamentally to agency (Sterelny 2001), given that the domains in which intelligence was selected for were predominantly practical: they concerned how organisms should avoid predators, find food, and so on. On this way of linking intelligence to agency, free actions might consist (at least partly) in the possession of features that enable human agents to behave intelligently in certain practical and moral domains.

Whatever the features are, many of them will likely favor the presence of (at least some of) the others by enhancing or supporting their functioning. In this way, the features will form a homeostatic cluster. Before returning to the question of which features, more finely specified, might belong in the cluster, it is useful to start by asking what functions are likely served by tracking any such cluster, and thereby what functions the concept FREE ACTION plausibly serves for us. Here, it is useful to begin by asking how we track agency in the first place, before we consider *free* agency.

Humans naturally distinguish agents from non-agents (e.g., Rutherford and Kuhlmeier 2013; cf. Heider and Simmel 1944; Bassili 1976).[12] Andrew Sims (2018) has argued that the concept AGENT is a natural-kind concept, which is subject to a

[12] There is some disagreement about whether detecting agency is strictly a matter of visual processing or is instead a matter of cognitive judgment. This dispute need not detain us in the present context.

non-descriptivist theory of reference. Sims appeals to *minimal mindreading* (Butterfill and Apperly 2013), which is the capacity to detect goal-directed behaviors by reliably detecting movements that function to bring about an outcome. Possession of a such a system puts one in an *epistemically rewarding* relation to other agents, by enabling one to gain useful information about them. In this way, Sims argues, we acquire a *proto*-concept of agents as a kind. This proto-concept lacks propositional or descriptive content, and so its reference works (and can only work) non-descriptively.

Over time, we add *beliefs* to the proto-concept, resulting in the mature concept, AGENT. While the beliefs have descriptive content, this content does not determine reference. There is, Sims maintains, "a continuity of reference between …[the]… proto-concept and the mature concept" (2018, p. 1), such that the mature concept also refers non-descriptively. In brief, Sims thinks that reference between the proto-concept and the mature concept must be continuous (and therefore non-descriptivist) because it is implausible to think that it would be entirely discontinuous, since then the proto-concept would be insufficiently related to the mature concept. Yet Sims also thinks it implausible that the continuity might be explained "in a 'satisfactional' way—that is, through the satisfaction of descriptional criteria" (2018, p. 8)—since the proto-concept appears *not* to be subject to any such requirement. As a result, AGENT is a natural-kind concept whose reference works non-descriptively, perhaps by means of our tracking a homeostatic cluster of properties (Sims 2018, p. 15).

Assuming roughly this view for AGENT, how might we acquire the concept FREE ACTION? The capacity to reliably detect agency is useful for us and feeds into other abilities, most notably causal cognition, which is the ability to predict

events, infer causes for events, and distinguish among types of causes. Tania Lombrozo (2010) argues convincingly that causal cognition works differently for people in different situations. When an event is categorized as occurring because of a transfer of energy, say, mechanistic thinking takes precedence. People judge the event as occurring because of the particular forces impacting on it (e.g., a tree falling in a storm). When an event is categorized instead as an action, people think more in terms of how counterfactually dependent its occurrence is on the goals of the agent whose action it was. For Lombrozo (2010, pp. 309–10), goal-directed agency exhibits *equifinality*—it results in a particular outcome (the one aimed at) across many different conditions. For example, Romeo's decision to reach Juliet will result in his doing so despite many possible obstacles. The behavior of non-goal-directed entities, by contrast, exhibits greater *multifinality*: different conditions result in different outcomes. For example, iron filings will not reach the magnet to which they are attracted even if a simple obstacle, such as a card, is placed between them.[13]

Of course, some events *not* categorized as actions may still exhibit equifinality to some degree. For example, a plant might grow around an obstacle blocking its sunlight. Such behavior is at least weakly equifinal, unlike the behavior of iron filings. Yet a plant's behavior is nowhere near as strongly goal-directed as Romeo's ability to overcome obstacles.[14]

Likewise, we are able to distinguish among *strengths* of equifinality. For example, a scrub jay's attempts to deceive its conspecifics (Bugnyar et al. 2016) will

[13] The examples of Romeo and the iron filings are drawn from William James (1890, p. 20).

[14] There is emerging evidence that plant behavior is far more strongly equifinal than we previously thought (Maher 2017). Similar findings have emerged regarding purposive agency in bacteria (Fulda 2017).

be categorized as far more weakly equifinal than Romeo's ability to reach Juliet.[15] Certain changes in background conditions will result in the scrub jay's failing to overcome an obstacle, but not Romeo. As Lombrozo puts it, "Goal-directed human behavior is a gold standard for equifinality" (2010, p. 310).

This ability to discriminate among types of (or degrees of sophistication in) agency plausibly extends to distinctions we make among human agents. It is presumably useful for us to learn to expect different behaviors from healthy fellow humans than, say, from schizophrenics (cf. Shoemaker 2015). These different expectations may result partly from our distinguishing the type of agency exercised by schizophrenics from the type exercised by "psychologically healthy, normally functioning adults" (McKenna 2012, p. 147). Above a certain threshold of sophistication in agency, which may go beyond the mere requirement that an action be intentionally performed, we start to categorize actions as *freely* performed.

Thus, FREE ACTION is plausibly a kind concept in folk psychology, understood as a set of cognitive abilities that enables us to make sense of others, including by predicting and explaining their behavior. Such prediction and explanation need not be explicit, since it might manifest simply as an implicit expectation about the relevant locus and type (or degree of sophistication) of control for behavior (or whatever other features might be required for equifinal human agency).[16]

[15] Deery and Nahmias (2017) explain our ability to distinguish among differing strengths of equifinality in terms of the strength of invariance between variables representing the output of an agent's cognitive processes and the event of the agent's obtaining a goal.

[16] In this way, on the HPC account it might additionally be an empirical and *a posteriori* theoretical question how we *track* whatever features underpin free actions (cf. Hutto and Myin 2017).

How might FREE ACTION relate to moral responsibility on this picture? For Vargas, FREE ACTION refers just in case it is suitably related to *something* that supports the central functional role of the concept, which he maintains is "to provide justified normative support for a sizable subset of our responsibility-characteristic practices, attitudes, and beliefs" (2006, pp. 353–4). Vargas maintains that FREE ACTION *is* suitably related to something that supports this role, namely, the capacity to respond to moral reasons, which is "cultivated" by the influence of our responsibility practices—as when we blame someone for failing to respond to moral reasons, or praise her for responding. Vargas insists that

> [T]his is not a picture according to which agents understand themselves to be trying to influence other agents. Instead, it is a two-tiered account, where the content of the norms ("act with due moral concern") make no appeal to effects, but the justification for continuing to enact those norms does appeal to the effects. …As such, it permits retrospective, desert-entailing… content in ordinary judgments. (2017, p. 231)

We need not adopt Vargas's proposal wholesale. Yet the agency-cultivation aspect of his view is attractive. Indeed, if FREE ACTION is a concept in folk psychology, then the mechanisms by which it contributes to agency-cultivation may be more numerous than Vargas realizes.

For example, defenders of *pluralistic folk psychology* (McGeer 2007; Andrews 2012; Zawidski 2013; cf. Spaulding 2018) maintain that belief-desire mindreading—typically considered as central to (or even definitional of) folk psychology—is merely

one of the capacities in our folk-psychological toolkit. These theorists suggest that even when humans do mindread by attributing beliefs or desires, their aim is not always to predict or explain behavior. Instead, it is to regulate behavior (McGeer 2007). Tad Zawidski (2013) maintains that folk psychology thus often aims at *mindshaping*, where the aim is to shape (not just read) other minds, such that they will conform to more predictable patterns of behavior, and in this way facilitate social cooperation. Zawidski further maintains that other tools apart from mindreading (in the traditional belief-desire sense) are central to folk psychology, yet do not always aim at prediction or explanation but instead at regulating others' behavior, even if this aim is not immediately apparent to us.

Within this framework, the idea that FREE ACTION serves normative functions seems hardly surprising. Furthermore, and by analogy with how the evolutionary niche occupied by a species results in selection pressure for specific traits, our folk psychological and moral responsibility practices plausibly function as extrinsic homeostatic mechanisms on a cluster of features in us, such that the practices themselves partly *determine* the reference of FREE ACTION.

That is enough for now to at least motivate the claim that FREE ACTION might be an HPC-kind concept in folk psychology. Furthermore, if free actions were not an HPC kind, it is difficult to see how we could ever have acquired the concept in the first place. Plausibly, we acquired it because there *is* a cluster of features that is useful for us to track, since doing so serves useful functions for us. In turn, our ability to track this cluster may itself be *part* of the cluster that defines free actions, since to predict, explain, or influence other agents' behavior successfully, sometimes

we must be able to track how well *they* can predict, explain, or influence *our* behavior (Sterelny 2003).

What other features might belong in the cluster? Ultimately, that is an empirical question. Recently, though, philosophers have proposed hypotheses that go beyond the stock suggestions that we be able to recognize and respond to reasons (Fischer and Ravizza 1998), including moral reasons (Wolf 1987), perhaps by reflecting on and identifying with our higher-order desires (Frankfurt 1971), while not acting on compulsive or compelled desires (Mele 1995). For instance, Eddy Nahmias (2018) suggests that the ability to *imagine* options and assess their likely outcomes may belong in the cluster that defines free actions. Relatedly, others have suggested that *prospection*, which is the mental (not necessarily conscious) simulation of future possibilities for the purpose of guiding action, also has a place in the cluster (Seligman et al. 2013; cf. Deery 2015b). These proposals also suggest that the cluster will be homeostatic, since how well or badly we exercise the ability to act in accordance with reasons (including moral reasons), for example, may depend on whether we possess and exercise other abilities, including some of the abilities just mentioned (cf. Nahmias 2018).

5. The Martian-control objection

Vargas notes that there has been "comparatively little direct uptake" (2011, p. 465) of placeholder preservationism in the literature. One reason is the Martian-control objection, which targets the suggestion that free actions might be a natural kind,

whose concept is amenable to a non-descriptivist semantics (Daw and Alter 2001; Balaguer 2010; cf. van Inwagen 1983).

Imagine we discover that paradigm cases of free action are Martian-controlled. According to Mark Balaguer (2010), who endorses the objection, on the Putnam-Kripke view "it would follow that free will consists in being controlled by Martians" (2010, p. 172). Yet that verdict may seem "absurd" (Daw and Alter 2001, p. 351).

This *Martian-control objection* fails, I maintain, even against the Putnam-Kripke view adopted by Heller (1996). According to Heller, the reference of FREE ACTION is "determined by paradigm cases, so that a …[free action]… is anything that is of the same kind as the paradigm cases" (1996, pp. 333–34). As Russell Daw and Torin Alter (2001) note in their assessment of Heller's view, "Being *the same kind* is a matter of sharing a certain underlying trait… which on the Putnam-Kripke view can be discovered only by empirical investigation" (p. 347). Along with causal-historical externalism, Heller adopts essentialism about the nature of the kind, such that "We have a particular action before us, and we ask 'what is essential to *this* kind of action?'" as a result of which "Discovering the essential nature of the kind becomes an empirical matter" (1996, p. 334).

Let us consider the objection, framed as an argument against the general view that free actions might be an empirically discoverable kind, and in particular against Heller's view:

(1) Because paradigm cases of free action are controlled by Martians, FREE ACTION does not refer (and free actions do not exist).[17]

(2) The natural-kind view entails that in this scenario, FREE ACTION refers (and so free actions exist).

Therefore,

(3) The natural-kind view about FREE ACTION (and free actions) is false.

In advancing this objection, Daw and Alter (2001) assume Heller's view in spelling out the "natural-kind view" in the second premise (2001, pp. 347–49). Nevertheless, they illicitly assume descriptivism in the first premise. Descriptivism says that FREE ACTION refers just in case no major presuppositions associated with the concept are in error. Yet, Daw and Alter suggest, surely a major presupposition about free actions *would* be in error were we to discover that the paradigm behaviors are Martian-controlled, since it might appear that an action's being free requires that it *not* be the result of manipulation, including of this sort, by another agent.[18] Daw and Alter think that in the situation as described, "the correct conclusion would… be that… no human actions are free" (2001, p. 351). In this way, they simply assume that we can answer important questions about the reference of FREE ACTION—and thus the existence of free actions—by reflecting on the intuitive presuppositions associated with the concept.

[17] Recall that "paradigm cases of free action" should be read as outlined in Section 3.

[18] In fact, it is a hotly debated issue in the free-will literature whether agents who are covertly manipulated can nevertheless act freely, as long as the manipulation unfolds through these agents' capacities to reason about what to do, reflectively endorse their actions, and so on (e.g., Pereboom 2014; McKenna 2014; Deery and Nahmias 2017).

Heller explicitly denies this assumption (1996, pp. 335–6). Consequently, Daw and Alter beg the question by formulating a descriptivist argument against a non-descriptivist treatment of FREE ACTION. They assume that descriptivism is the correct theory.

By contrast, if we apply the Putnam-Kripke view, as Heller does, then there is no reason whatsoever to believe premise (1). As Heller puts it, were we to discover that what "paradigms of free action" (1996, p. 335) have *essentially* in common is that they are controlled by Martians, then on the Putnam-Kripke view, "We cannot find that it is essential to actions of this kind that they not be caused by desires caused by …[Martians]…, because it has turned out that these actions are caused in that way" (1996, p. 334). For Heller, reference is fixed by paradigm cases, such that the concept refers to all and only actions sharing those cases' essential feature. If that feature is Martian control, so be it.

As a result, the Martian-control objection fails, even against Heller. Nevertheless, if what "paradigms of free action" (Heller 1996, p. 335) have essentially in common is that they are Martian-controlled, then according to Heller their being controlled in this way is what *makes* them free. By contrast, the HPC view avoids this counterintuitive result by placing additional (independently motivated) constraints on what counts as a free action, such that FREE ACTION cannot refer because of Martian-control.

6. Martian-control on the HPC view

Assume that paradigm cases of the most sophisticated agency typically exercised by humans in certain contexts are actually Martian-controlled behaviors (if it helps,

assume that *all* human behaviors are Martian-controlled). Now grant that FREE ACTION has something like a stabilizing function, which works to support agreement about the cases to which the concept applies. There could be such a function only if the Martian control is subtle, in not violating any of the agential features that regulate our acquisition and use of the concept. If it *did* violate those features, then either we would not have acquired the concept or it would not perform the work that we require of it. Assuming the concept *has* a stabilizing function (as it appears to), the Martians' control must be subtle.

Here is one way it might be subtle. Imagine that the homeostatic cluster of features the paradigm cases have in common, and which exerts a stabilizing function on our acquisition and use of the concept, is located in us, in humans. Yet this cluster's homeostasis is maintained centrally by an extrinsic mechanism operated by the Martians, who intend for us to perform certain actions. Our actions are Martian-controlled by means of this extrinsic mechanism, despite their unfolding via a cluster of agential features located in us.[19]

Now recall the two premises of the Martian-control objection:

(1) Because paradigm cases of free action are controlled by Martians, FREE ACTION does not refer (and free actions do not exist.)

(2) The natural-kind view entails that in this scenario, FREE ACTION refers (and so free actions exist).

[19] If the Martians control only *some* of our actions in this way, then the following considerations apply only to those actions. I thank a referee for prompting me to note this consequence of the view.

Once we fill in the details of the "natural-kind view" in premise (2) with the HPC theory, there is no way for premises (1) and (2) to be true together. Either (1) is false because (2) is true, or (2) is false, but not because of Martian control (and thus (1) is false too).

Consider the first possibility. For FREE ACTION to have a stabilizing function, some cluster of features must do a good enough job of serving our goals in tracking it. So the Martian control must be extremely subtle. As suggested, it must work as something like an extrinsic homeostatic mechanism that maintains a cluster of agential features in us, such that we perform the actions that the Martians want, yet we do not need to identify this mechanism to at least apparently track the locus and type of control for one another's behaviors (or to track whatever other features may be required for the most sophisticated type of agency typically exercised by humans). Thus, a cluster we already track seems sufficient for FREE ACTION to refer given our goals in using it, yet not because of the unifying role of Martian control. So premise (1) is false, because (2) is true.

This verdict is analogous to a well-worn compatibilist response to *manipulation arguments* for the incompatibility of free will and determinism (e.g., Pereboom 2014; Mele 2013). Such arguments claim, first, that as a result of their being covertly (and deterministically) manipulated, agents do not act freely, even if they satisfy any combination of compatibilist conditions proposed as sufficient for free action (e.g., Frankfurt 1971; Fischer and Ravizza 1998; Wolf 1987; Mele 2006). Second, these arguments claim that there is no principled difference between manipulation of this sort and ordinary causal determinism when it comes to acting freely. The conclusion

drawn is that determined agents do not act freely either, and thus free actions are incompatible with determinism.

Michael McKenna (2008) calls compatibilist responses that reject the first claim—that manipulated agents do not act freely—*hard-line* responses. He calls compatibilist responses that reject the second claim—that there is no relevant difference between manipulation and determinism—*soft-line* responses (see also Kane 1996).

The analog hard-line in response to the Martian-control objection says that if most of our goals (or our most important goals) are served by the homeostatic cluster we already track, then FREE ACTION refers—even in the Martian-control scenario. Yet it does not refer *because* of Martian control, as on Heller's view. Instead, it refers *notwithstanding* such control. So free actions exist, despite the Martians' control. On the analog hard-line, premise (2) of the Martian-control objection might be true, but (1) is false because (2) is true. Advocates of the objection must assume (question-beggingly) descriptivism for (1) to be true. Assuming the HPC view instead, the objection fails because (1) is false.

It is *possible* on Heller's view that the concept FREE ACTION might not refer—for instance, if the paradigm cases lack any unifying essence at all and we "just accidentally latched onto a coincidental conjunction of unrelated features, an arbitrary hodge-podge" (Hurley 2000, p. 234). Yet that is not the case in the Martian-control scenario, since on Heller's view Martian control *is* (counterintuitively) the essence that the paradigm cases have in common (Heller 1996, p. 335).

Likewise, the HPC view allows for the possibility that FREE ACTION might not refer. Yet unlike Heller's view, it also allows for this possibility in the Martian-

control scenario (depending on the details of how the HPC cluster and the manipulation are specified). In that case, premise (2) would be false—the concept FREE ACTION would not refer. This verdict is analogous to soft-line responses to manipulation arguments, which reject the claim that there is no relevant difference between manipulation and determinism when it comes to acting freely. Soft-liners grant that because of the way in which an agent might be manipulated, she does not act freely; yet they deny that determinism similarly undermines free action. One recent soft-line position claims that agents who are covertly manipulated do not act freely because the causal sources of their actions lie beyond themselves in the intentions of the manipulator, which is not true of merely determined agents (Deery and Nahmias 2017).

In a similar vein, discovering the Martians' control might lead us to decide that even if FREE ACTION had functioned well enough for us, our newfound knowledge should make us look elsewhere for the locus of control for the relevant behaviors, and to downgrade the type of control that we attribute to the agents in question. We would look for a causal source for these behaviors beyond the agents themselves. Alternatively, it might be that other goals of a pluralistic folk-psychology (such as those hinted at in Section 4) would no longer be adequately served, with the result that reference would be similarly undermined. Either way, (2) would be false—the "natural-kind view" would *not* entail that FREE ACTION refers in this scenario. That result is enough to show that the Martian-control objection fails.

Consequently, we avoid the counterintuitive result of Heller's view that FREE ACTION might refer because of Martian control. On the HPC view, even if the

concept fails to refer due to Martian control, clearly it does not refer *because* of Martian control, since the concept fails to refer at all.

For present purposes, I need not decide whether to adopt the HPC analog soft-line or instead the hard-line response to the Martian-control objection, since it is enough to show that the objection fails either way. On both options, the HPC view avoids the result that FREE ACTION refers because of the unifying role played by Martian control, which is a bullet that Heller must (and does) bite. The HPC view is thus able to avoid being counterintuitive in this way, for whatever that is worth.

One thing that avoiding this counterintuitiveness may buy the HPC view is a dialectical advantage in engaging with descriptivists, for whom appeal to intuitions is paramount in formulating a theory of free action. To the extent that the HPC analog hard-line requires biting any counterintuitive bullets at all, matters are no worse than they are for hard-line compatibilists such as McKenna. Either the Martians' manipulation works while keeping the manipulated agents' ordinary agentive features intact, in which case these agents act freely notwithstanding the Martians' control, or the manipulation fails to keep the features intact, in which case the "jig is up," as McKenna puts it (2008, p. 144), and the agents do not act freely.

Additionally, the HPC view has further advantages over Heller's view, as we shall now see. Due to these advantages, the HPC view should be preferred to Heller's view as the default natural-kind or preservationist view to hold about free actions.

7. Metaphysics and semantics (again)

Heller's adoption of essentialism about the metaphysics of kinds is highly controversial, at least among naturalistic philosophers (e.g., Dupré 1981; Sterelny and

Griffiths 1999). As we have seen, the HPC view avoids this commitment to essentialism by relying instead on the idea that kinds are homeostatic property clusters. In this way, my view avoids the metaphysical difficulties that Heller's view inherits. Yet the HPC view also avoids another problem Heller inherits due to his adopting causal-historical externalism.

According to causal-historical externalism, the reference of a concept like WATER, for example, is determined by perceptual contact with a sample of the kind together with the assumption that the term or concept picks out *the same liquid* as the sample. However, as Millikan notes, this procedure trades on our thinking that there are "special … rules we have all learned for determining that 'water' must always name 'the same liquid' and legislating some common way that we should all understand 'same liquid'" (2010, p. 65). Instead, Millikan maintains that what supports agreement in people's judgments "is not some criterial way everyone has been taught for recognizing water"; instead, "it is the contingent but lawful clustering of distinctive traits of water all produced by the same molecular structure" (2010, p. 65).

According to Millikan, the causal-historical view illicitly incorporates a descriptivist element by relying on a "criterial" way of recognizing "the same kind" (e.g., for water: *the odorless, clear liquid that fills lakes, flows in rivers, falls as rain*, etc.). This procedure enables us to generalize from the paradigm cases to the other members of the kind. Yet the procedure makes the causal-historical view unstable. Either we have

no way of generalizing from the paradigm cases or else the causal-historical view collapses into descriptivism, to which it was supposed to offer an alternative.[20]

Instead, we should think of what supports agreement in our judgments and enables us to generalize from the paradigm cases as being a lawful clustering of homeostatic features (Millikan 2010, pp. 59–66). That is, we acquire and retain a concept like WATER due to its "stabilizing function" (Millikan 2010, pp. 43–81), which is simply its survival value or usefulness for us. Like the Putnam-Kripke view, this view is externalist (and thus non-descriptivist) since the meaning of WATER depends on whether deployments of the concept share a stabilizing function that results from a homeostatic clustering of features in the world. Yet in contrast to how reference works on the causal-historical view, here "reference… is entirely *direct* … It has no defining intension" (Millikan 2010, p. 65).

These considerations apply, *mutatis mutandis*, to FREE ACTION. As a result, Heller's view either has no way of generalizing from the paradigm cases or it collapses into descriptivism. My HPC view avoids this difficulty by relying on something like Millikan's idea of a stabilizing function, according to which a contingent but lawful clustering of features in the world regulates our acquisition, retention, and continued use of the concept FREE ACTION, which in turn supports agreement about the cases to which it applies (as I pointed out in Section 3).

Accordingly, we identify the paradigm cases not by asking whether we would explicitly apply the concept FREE ACTION to a concrete case (much less to an exotic, merely possible case) on the basis of some criterial or descriptive way of

[20] For further difficulties with this procedure, see Laurence and Margolis (2003).

identifying such actions. Instead, we acquire the concept and identify the paradigm cases by means of our ability to track a stable cluster of features in other agents, due to the usefulness of our being able to do so, which in turn suggests that there *is* a cluster of features and mechanisms that we keep track of, even if we must empirically discover what they are. Such an account is a variety of placeholder preservationism, since it endorses a default realism about the existence of free actions.

Notice that this account brings *us* into the picture in a way that causal-historical externalism does not. A cluster of features in agents causes a stabilizing of the concept in us, but only because of the usefulness of our tracking that cluster. A consequence of this aspect of the view is that what *fixes* reference for FREE ACTION is partly condition (b), as stated in Section 3—i.e., the requirement that the cluster serve our goals in tracking it. The metaphysical and semantic stories simply do not come apart as easily on the HPC view as they do on the Putnam-Kripke view, once we take seriously the requirement that a kind concept have a stabilizing function.

The HPC view both avoids Heller's commitment to essentialism and explains how we generalize from the paradigm cases (or identify them in the first place), yet without illicitly assuming descriptivism.

8. Clarifications and replies to objections

8.1 Is the HPC view revisionist?

Vargas maintains that reference for FREE ACTION might be secured despite errors associated with the concept, by abandoning the erroneous presuppositions (2011, p. 460), a view he calls revisionism. Is the HPC view revisionist? I maintain that the

HPC view is compatible with revisionist and non-revisionist views about the natural kind, but that unlike Vargas's view, it need not make a bet in advance about whether it is revisionist.

Vargas tries to remain agnostic on the details of how reference works (2011, p. 468). If an erroneous presupposition is reference-fixing, and thus reference fails, Vargas thinks that we might re-anchor reference to something that preserves the central role of the concept, an option he calls "denotational revisionism." While some revisionists endorse this view (e.g., McCormick 2016), Vargas worries that it collapses into eliminativism, since it starts out by granting reference failure.[21]

By contrast, if an error is not reference-fixing, reference might succeed in spite of the error, an option Vargas calls "connotational revisionism." For instance, we might accept, with Frank Jackson, that FREE ACTION "embodies some kind of confusion" (Jackson 2001a, p. 618), yet insist that an account can be provided of how reference is secured despite this error (2001b, p. 661). In this way, Vargas could be interpreted as adopting a liberal *descriptivist* approach (i.e., one that is tolerant of at least some errors). In places, however, Vargas suggests that he may be open to a non-descriptivist variety of revisionism, by accepting a type of natural-kind view (2013, p. 95).

I deny that a natural kind view must be revisionist, at least according to Vargas's own characterization of such a position. Although it is obvious that any natural-kind view, including my HPC view, permits (even radical) revision away from

[21] McCormick (forthcoming) argues that there is a better case to be made for reference success than reference failure, even for denotational revisionists.

whatever commonsense presuppositions might be associated with the concept, it does not *require* such revisions, whereas Vargas's view does.

In his most recent statement of the view, Vargas characterizes revisionism as follows: "What makes a theory revisionist… is not merely… departure from commonsense. Instead, it is the fact that the truth of the proposal is in *conflict* with commonsense" (forthcoming). In other words, a theory is revisionist only if it conflicts with commonsense presuppositions associated with the concept; thus, conflict is required. However, because the HPC view makes it an open empirical question what features (if any) might underpin free actions, strictly speaking it leaves it open that commonsense presuppositions might be *true*. Each presupposition about free actions, and each set of (perhaps settled) presuppositions (e.g., compatibilist or libertarian), amounts on the HPC view to a hypothesis about what free actions consist in. These hypotheses must be empirically tested. If a presupposition (or set of them) is confirmed, revisionism is false; yet a natural-kind view (like the HPC view) might well be true. As a result, the HPC view need not be revisionist, by Vargas's own lights.

8.2 The intuition objection

One might object that the HPC view relies on intuitions just as much as descriptivist views in picking out the paradigm cases. I deny this claim (see sections 3 and 7, above). Yet even if it were true, it poses no difficulty for the HPC view. After all, even for non-descriptivists intuitions need not be philosophically worthless.

For one thing, intuitions presumably reflect our beliefs and presuppositions about a kind, and so they can help us to orient (especially initial) investigation into a

kind. It does not follow, however, that intuitions about a kind are not revisable upon subsequent findings. Moreover, as Stephen Laurence and Eric Margolis put it, "Intuitions can provide evidence about the content of a concept. …[and]… They can do this even if they aren't constitutive of conceptual content, since intuitions may nonetheless be correlated with the conditions that are" (2003, p. 278). For Laurence and Margolis, intuitions reflect our dispositions to categorize, which typically correlate (non-accidentally) with conceptual content. As a result, "*intuitions are broadly correlated with content on virtually any theory of content* [italics in original]" (2003, p. 279). Thus, intuitions "can be used as evidence for content, that is, evidence that a concept truly applies when intuition says it does" (p. 279). This much might be true even if intuitions are not constitutive of conceptual content.

On the HPC view, FREE ACTION refers (if it does) to those behaviors that we have been trying all along to pick out by means of our intuitive presuppositions. Even so, reference does not depend on whether the content of these presuppositions is satisfied by what is actually referred to. Instead, it depends on HPC criteria. If the concept does refer, our presuppositions (even if erroneous) are still *about* its actual referent.

8.3 The false-belief objection

Relatedly, one might worry that the falsity of (at least some of) our central presuppositions about free actions might undermine (at least some of) our moral practices, at least if these practices are based on the presuppositions in question. For example, if retributive punishment presupposes that agents exercise indeterminist free will, yet we have reason to believe that no one has such free will, our practice of

punishing might be put into doubt (Caruso 2012; cf. Singer 2002). This outcome remains possible on the HPC view. However, it is likely that any such impact would be slight, and would not undermine many of the most central goals served by FREE ACTION, since the concept appears to have a stabilizing function, which implies that there *is* a cluster of features and mechanisms that we track, since our doing so serves (at least many of) the goals that we require our tracking it to serve.

8.4 The kind objection

Perhaps free actions are merely a social kind. To the extent that reference is determined on the HPC view by how our tracking a cluster serves goals for us, if these goals (or the features in the cluster) differ across communities, then there may be no *universal* or "natural" kind picked out by the concept. According to Tamler Sommers (2012), for instance, some cultures justify praise, blame, and punishment even when agents lack the sort of features required by philosophers for free action. What Sommers calls "honor cultures" condone punishing agents even for their relatives' transgressions, whereas non-honor cultures like our own condemn such punishments as unjust. For Sommers, cultural variations of this sort result from different modes of life and adaptations to different conditions. As a result, cultures might track different clusters to serve similar goals to ours, or serve different goals than we do by tracking the same cluster we track.

This objection poses no difficulty for the HPC view. For one thing, the distinction between natural kinds and social kinds is not as sharp on the HPC view as on other views. Even if the HPC view does not "carve nature at its joints" in identifying universal kinds, it nevertheless *carves nature into useful chunks*. This aspect of

the view is unavoidable, since the existence of a natural category or kind depends, according to the HPC view, on the goals served by our tracking a stable cluster of features. Even so, such kinds are not nominal, since they are tied to causal structure. If different communities track different clusters than we do, or if they track similar clusters to serve different goals than we do, the kinds in question may well be different. After all, the reality of any HPC kind is not assessed apart from the role it plays in a domain of practices: if it plays an appropriate role, it is real. Consequently, while the concept of a lily does not normally apply to onions or garlic, the botanical kind *Liliaceae*, to which the flowers we call lilies belong, actually includes these vegetables, as well as various tulips (Dupré 1981). Thus, lilies—as a kind from ordinary life—are distinct from the botanical kind. Yet lilies are plausibly an HPC natural kind (Boyd 1999, p. 161). After all, they share a homeostatic cluster of aesthetic properties (e.g., of coloration and structure) that enables horticulturalists and gardeners to achieve certain goals precisely because the cluster accommodates the demands of these practical domains by achieving accommodation to causal structure.

As a result, it might be that different cultures have different, perhaps related, concepts of free action, and thus correspondingly different kinds exist. Perhaps *we* are interested mostly in our own local concept—as we should be, since it is important. Yet having that interest is consistent with our investigating other kinds in different cultures, and comparing how these potentially related kinds function.

8.5 The control objection

The focus in free-will debates is often on explaining free actions in terms of an agent's exercising the strongest type of control required for moral responsibility (e.g., McKenna 2012, p. 17). Numerous accounts take this form (e.g., Kane 1996; Fischer and Ravizza 1998; Clarke 2003; Mele 2006). Sometimes, the relevant control is understood as including the ability to do otherwise (van Inwagen 1983; cf. Vihvelin 2004; Ismael 2013), while at other times what is considered more important is whether an agent is the relevant *source* of her action, even if she could not have done otherwise (Frankfurt 1969; Fischer and Ravizza; Mele 2006; Pereboom 2014). Other theorists, however, deny that control is required. They maintain instead that for an agent's acting freely her action must be suitably attributable to her, perhaps because it is caused by attitudes that reflect her "real self" or character (e.g., Smith 2008), or because the action expresses the agent's quality of regard for the moral standing of others (e.g., Arpaly 2003; cf. Shoemaker 2015). Control is not required. The objection is that the HPC view fails to take these alternatives into account.

For simplicity's sake, let us stipulate that all of the accounts just mentioned aim at explaining free actions, on the assumption that these are the actions (if any) for which agents are morally responsible. Accordingly, some philosophers explain free actions in terms of control, while others do so in terms of quality of regard, and so on. All of these approaches are consistent with the HPC view that I have outlined above, since it maintains that what explains free actions is an empirical and *a posteriori* theoretical matter. In other words, even though I have occasionally written in terms of control, strictly speaking the HPC view is specified in terms of *features* of agents,

which may include control capacities but also other features, such as an agent's quality of regard for others, and so on.

Relatedly, one may object that my occasional focus on agents' being the *locus* of control is not enough, since we must distinguish mere intentional behavior (for which agents may be the locus of control) from *free* actions. Recall, however, that the HPC view focuses not merely on the locus but also the *type* (or degree of sophistication) of control for intentional actions (insofar as it focuses on control at all), as outlined in Section 4. The view thus permits testing for most of the hypotheses suggested by the going theories of free will, whether libertarian (e.g., Kane 1996) or compatibilist.

11. Concluding remarks

Against the odds, a lacuna remained in the logical geography of theorizing about free actions. The HPC view fills out this uncharted space. Yet the view does more than simply map out philosophical terrain. It provides a new way of thinking about free actions, which has explanatory power and excels where other theories fall short. Free actions are a plausibly an HPC kind, and we humans act freely, at least sometimes, as long as we possess various features that are related in the right sorts of ways to each other and to the world. In turn, we acquire and retain the concept FREE ACTION as long as most of us possess enough of these features. Our presuppositions about free actions, whatever they happen to be, are beside the point when it comes to whether FREE ACTION refers, or free actions exist. The default answer is that it does, and they do.

Acknowledgments Versions of this paper were presented at the University of Arizona (January 2016), Florida State University (January 2017), Monash University (September 2017), the Central Division Meeting of the American Philosophical Association (February 2018), the University of Melbourne (May 2018), and the 3rd International Conference on Natural Cognition at the University of Macau (November 2018). Thanks to audiences at those venues for helpful comments. I also thank Terry Horgan, Shaun Nichols, Michael McKenna, Eddy Nahmias, Alfred Mele, Tim Bayne, Gregg Caruso and several anonymous referees for their valuable suggestions. Finally, I thank the students in my seminar on reference and free will at Monash University (2018) for useful discussion.

References

Andrews, K. (2012). *Do apes read minds? Toward a new folk psychology*. Cambridge, MA: The MIT Press.

Arpaly, N. (2003). *Unprincipled virtue: An inquiry into moral agency*. Oxford: Oxford University Press.

Balaguer, M. (2010). *Free will as an open scientific problem*. Cambridge, MA: The MIT Press.

Bassili, J. (1976). Temporal and spatial contingencies in the perception of social events. *Journal of Personality and Social Psychology*, *33*(6): 680–685.

Boyd, R. (1988). How to be a moral realist. In G. Sayre-McCord (Ed.), *Essays on moral realism* (pp. 181–228). Ithaca, NY: Cornell University Press.

Boyd, R. (1999). Homeostasis, species, and higher taxa. In R. A. Wilson (Ed.), *Species: New interdisciplinary essays* (pp. 141–85). Cambridge, MA: The MIT Press.

Brigandt, I. (2011). Natural kinds and concepts: A pragmatist and methodologically naturalistic account. In J. Knowles and H. Rydenfelt (Eds.), *Pragmatism, science and naturalism* (pp. 171–96). Frankfurt am Main: Peter Lang Publishing.

Bugnyar, T., Stephen A. Reber, S. A., & Buckner, C. (2016). Ravens attribute visual access to unseen competitors. *Nature Communications*, *7*, 10506.

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind and Language*, *28*: 606–637.

Caruso, G. (2012). *Free will and consciousness: A determinist account of the illusion of free will.* Lanham, MD: Lexington Books.

Caruso, G. (2015). Free will eliminativism: Reference, error, and phenomenology. *Philosophical Studies*, *172*(10): 2823–2833.

Clarke, R. (2003). *Libertarian accounts of free will.* New York: Oxford University Press.

Daw, R., & Alter, T. (2001). Free acts and robot cats. *Philosophical Studies*, *102*: 345–357.

Deery, O. (2015a). The fall from Eden: Why libertarianism isn't justified by experience. *Australasian Journal of Philosophy*, *93*(2): 319–334.

Deery, O. (2015b). Why people believe in indeterminist free will. *Philosophical Studies*, *172*(8): 2033–2054.

Deery, O., & Nahmias, E. (2017). Defeating manipulation arguments: Interventionist causation and compatibilist sourcehood. *Philosophical Studies*, *174*(5): 1255–1276.

Deery, O., Davis, T., & Carey, J. (2015). The Free-Will Intuitions Scale and the question of natural compatibilism. *Philosophical Psychology*, *28*(6): 776–801.

Dupré, J. (1981). Natural kinds and biological taxa. *The Philosophical Review*, *90*(1): 66–90.

Ereshefsky, M., & Reydon, T. A. C. (2015). Scientific kinds. *Philosophical Studies*, *172*(4): 969–986.

Fischer, J., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.

Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy*, *66*(23): 829–839.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, *68*(1): 5–20.

Fulda, F. C. (2017). Natural agency: The case of bacterial cognition. *Journal of the American Philosophical Association*, *3*(1): 69–90.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*(2): 243–259.

Heller, M. (1996). The mad scientist meets the robot cats: Compatibilism, kinds, and counterexamples. *Philosophy and Phenomenological Research*, *56*(2): 333–337.

Hurley, S. (2000). Is responsibility essentially impossible? *Philosophical Studies*, *99*(2): 229–268.

Hutto, D., & Myin, E. (2017). *Radicalizing enactivism: Basic minds without content*. Cambridge, MA: The MIT Press.

Ismael, J. (2013). Causation, free will, and naturalism. In H. Kincaid, J. Ladyman & D. Ross (Eds.), *Scientific metaphysics* (pp. 208–235). New York: Oxford University Press.

Jackson, F. (1998). *From metaphysics to ethics: A defence of conceptual analysis*. Oxford: Oxford University Press.

Jackson, F. (2001a). Précis of From metaphysics to ethics. *Philosophy and Phenomenological Research*, *62*(3): 617–624.

Jackson, F. (2001b). Responses. *Philosophy and Phenomenological Research*, *62*(3): 653–664.

James, W. (1890). *The principles of psychology*. Cambridge, MA: Harvard University Press.

Kane, R. (1996). *The significance of free will*. New York: Oxford University Press.

Khalidi, M. A. (2018). Natural kinds as nodes in causal networks. *Synthese*, *195*(4): 1379–1396.

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

Kumar, V. (2014). 'Knowledge' as a natural kind term. *Synthese*, *191*: 439–457.

Kumar, V. (2015). Moral judgment as a natural kind. *Philosophical Studies*, *172*(11): 2887–2810.

Laurence, S., & Margolis, E. (2003). Concepts and conceptual analysis. *Philosophy and Phenomenological Research*, *67*(2): 253–282.

Levy, N. (2011) *Hard luck: How luck undermines free will and moral responsibility*. Oxford: Oxford University Press.

Levy, N. (2016). Implicit bias and moral responsibility: Probing the data. *Philosophy and Phenomenological Research*, *94*(1): 3–26.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, *50*(3): 249–258.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4): 303–332.

Maher, C. (2017). *Plant minds: A philosophical defense*. New York: Routledge.

Mallon, R., Machery, E., Nichols, S., & Stich, S. (2009). Against arguments from reference. *Philosophy and Phenomenological Research*, *79*(2): 332–356.

May, J. (2014). On the very concept of free will. *Synthese*, *191*(12): 2849–2866.

McCormick, K. A. (2016). Revisionism. In K. Timpe, M. Griffith & N. Levy (Eds.), *Routledge companion to free will* (pp. 109–120). New York: Routledge.

McCormick, K. A. (forthcoming). Meeting the eliminativist burden. *Social Philosophy & Policy*, *36*(1).

McGeer, V. (2007). The regulative dimension of folk psychology. In D. Hutto & M. Ratcliffe (Eds.), *Folk psychology reassessed* (pp. 137–56). Kluwer/ Springer Press.

McKenna, M. (2008). A hard-line reply to Pereboom's four-case manipulation argument. *Philosophy and Phenomenological Research*, *77*(1): 142–159.

McKenna, M. (2012). Moral responsibility, manipulation arguments, and history: Assessing the resilience of nonhistorical compatibilism. *Journal of Ethics*, *16*: 145–174.

McKenna, M. (2014). Resisting the manipulation argument: A hard-liner takes it on the chin. *Philosophy and Phenomenological Research*, *89*(2): 467–484.

Mele, A. (1995). *Autonomous agents: From self-control to autonomy*. New York: Oxford University Press.

Mele, A. (2006). *Free will and luck*. New York: Oxford University Press.

Mele, A. (2013). Manipulation, moral responsibility, and bullet biting. *Journal of Ethics*, *17*(3): 167–184.

Millikan, R. G. (2010). On knowing the meaning; with a coda on Swampman. *Mind*, *119*(473): 43–81.

Nahmias, E. (2018). Free will as a psychological accomplishment. In D. Schmidtz & C. E. Pavel (Eds.), *The Oxford handbook of freedom* (pp. 492–507). New York: Oxford University Press.

Nichols, S. (2015). Free will and error. In S. Nichols, *Bound: Essays on free will and responsibility* (pp. 54–74). Oxford: Oxford University Press.

Nichols, S. (Forthcoming). Free will and reference. In J. Campbell (Ed.), *A companion to free will*. Hoboken, NJ: Wiley-Blackwell.

Pereboom, D. (2014). *Free will, agency, and meaning in life*. Oxford: Oxford University Press.

Pober, J. M. (2013). Addiction is not a natural kind. *Frontiers in Psychiatry*, *4*: 123.

Putnam, H. (1975). The meaning of "Meaning." *Minnesota Studies in the Philosophy of Science*, *7*: 131–193.

Reydon, T. A. C. (2009). How to fix kind membership: A problem for HPC theory and a solution. *Philosophy of Science*, *76*(5): 724–736.

Rutherford, M. D., & Kuhlmeier, V. A., Eds. (2013). *Social perception: Detection and interpretation of animacy, agency, and intention*. Cambridge, MA: The MIT Press.

Salmon, N. (1982). *Reference and Essence*. Oxford: Basil Blackwell.

Seligman, M., Railton, P., Baumeister, R., & Sripada, C. (2013). Navigating into the future or driven by the past: Prospection as an organizing principle of mind. *Perspectives on Psychological Science*, *8*(2): 119–141.

Schechtman, M. (2014). *Staying alive: Personal identity, practical concerns, and the unity of a life*. Oxford: Oxford University Press.

Shoemaker, D. (2015). *Responsibility from the margins*. New York: Oxford University Press.

Singer, I. (2002). Freedom and revision. *Southwest Philosophy Review*, *18*(2): 25–44.

Sims, A. (2018). The essence of agency is discovered, not defined: A minimal mindreading argument. *Philosophical Studies*. doi: 10.1007/s11098-018-1108-5

Smith, A. (2008). Control, responsibility, and moral assessment. *Philosophical Studies*, *138*(3): 367–392.

Sommers, T. (2012). *Relative justice: Cultural diversity, free will, and moral responsibility*. Princeton University Press.

Spaulding, S. (2018). *How we understand others: Philosophy and social cognition*. New York: Routledge.

Sterelny, K. (2001). *The evolution of agency and other essays*. Cambridge: Cambridge University Press.

Sterelny, K. (2003). *Thought in a hostile world: The evolution of human cognition*. Oxford: Blackwell Publishing.

Sterelny, K., & Griffiths, P. (1999). *Sex and death*. Chicago: University of Chicago Press.

Strawson, G. (1986). *Freedom and belief*. Oxford: Oxford University Press.

Van Inwagen, P. (1983). *An essay on free will*. Oxford: Oxford University Press.

Vargas, M. (2006). On the importance of history for responsible agency. *Philosophical Studies*, *127*: 351–382.

Vargas, M. (2011). Revisionist accounts of free will: Origins, varieties, and challenges. In R. Kane (Ed.), *The Oxford handbook of free will, 2nd edition* (pp. 457–84). New York: Oxford University Press.

Vargas, M. (2013). *Building better beings: A theory of moral responsibility*. New York: Oxford University Press.

Vargas, M. (2017). Implicit bias, responsibility, and moral ecology. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility, volume 4* (pp. 219–247). Oxford: Oxford University Press.

Vargas, M. (forthcoming). Revisionism. In J. Campbell (Ed.), *A companion to free will*. Hoboken, NJ: Wiley-Blackwell.

Vihvelin, K. (2004). Free will demystified: A dispositional account. *Philosophical Topics*, *32*(1–2): 427–450.

Wilson, R. A., Barker, M. J., and Brigandt, I. (2007). When traditional essentialism fails: Biological natural kinds. *Philosophical Topics*, *35*: 189–215.

Wolf, S. (1987). Sanity and the metaphysics of responsibility. In F. Schoeman (Ed.), *Responsibility, character, and the emotions: New essays in moral psychology* (pp. 46–62). Cambridge: Cambridge University Press.

Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition*. Cambridge, Mass: Bradford Books.