

Can We Have Free Will and Lack It Too?

Joshua May

Originally published on the [Experimental Philosophy blog](#) (08/13/2015).

This summer I read [Shaun Nichols](#)'s excellent new book, *[Bound: Essays on Free Will & Responsibility](#)* (2015, OUP). It includes systematic discussion of the relevant experimental philosophy, as well other empirical research, and how this relates to the problem of free will and related topics (including responsibility and punishment).

I wrote a short (400-word) book note on *Bound* for the *AJP* ([draft here](#)), which prompted my further thinking and motivated a blog post. You can read a more detailed and comprehensive summary [by Manuel Vargas in NDPR](#) that raises some nice worries for Shaun. Here I'd like to summarize just one key component of *Bound* and raise a distinct worry for it (despite agreeing with many other claims made in the book).

This key component is Nichols's *discretionary view*. It states roughly that, assuming our actions probably are determined by the past and laws of nature, *there are contexts in which it's true that we have free will but also contexts in which it's false*. Can we really have our cake and *lack* it too?

Discretionary (In)compatibilism

Nichols argues that we can be drawn to the idea that freedom and responsibility are compatible with one's choices being determined by the past and laws of nature because we naturally seek deterministic explanations of everything. However, we develop our notion of choice in what we assume is an indeterministic world, because we implicitly think (erroneously) that we'd be able to tell if our decisions were determined. Hence, we implicitly assume that [determinism](#) undermines freedom and responsibility (contra [Nahmias and company](#)), but we're also drawn to the idea that it might not.

Our actions may well be determined, says Nichols, but we needn't deny that we're free and responsible, despite an implicit commitment to incompatibilism in ordinary discourse. Sometimes to successfully refer to these phenomena we must be talking about an undetermined ability to choose. But free will is like other entities—e.g. memory and solidity—that we can sensibly believe in even if our folk theories about them are partly in error.

The key example is of [catoblepas](#), which were once thought to be a distinct species but were probably based on sightings of wildebeests. Naturally, participants in [an experiment](#) disagree with "Catoblepas exist," presumably via a [descriptivist](#) convention of reference (e.g. catoblepas are the unique animals with such-and-such features). However,



people tend to *agree* with "Catablepas are wildebeests," even though the statement assumes existence. Here participants are presumably using a causal-historical convention of reference (e.g. catoblepas are the animals people encountered and dubbed "catoblepas"). This isn't a matter of confusion (and *perhaps* not a mere verbal dispute?). False beliefs about a topic can generate different conditions of reference in different contexts.

Similarly, according to Nichols, there is a sense in which we lack free will (eliminativism) and a sense in which it's just not quite what we thought it was (revisionism). We encounter in ourselves and others an ability to make choices and control our behavior based on our preferences, dispositions, reasoning, etc. We can refer to this ability descriptively—as something that's incompatible with determinism—in which case Nichols thinks we (probably) lack such an ability. But we can also refer to the ability causal-historically: as the ability we encounter in us and call "free will," which grounds moral responsibility. Using this convention for reference, Nichols thinks we do have this ability, since incompatibilist commitments are "intuitive but isolated"—they can be jettisoned while largely preserving our practices of praise and blame.

A Worry: Arbitrariness

This "discretionary (in)compatibilism" strikes me as *problematically arbitrary*, as when it apparently implies that one can give up the incompatibilist commitment when it's prudent to do so. For instance, as a therapeutic response to overwhelming guilt or shame (p. 166), one might take comfort in the idea that "free will doesn't exist." And the statement would apparently be true!

Yet it doesn't seem to go this way in the other cases of shifting reference conditions on which free will is modeled. For "memory" or the curious "catoblepas," there seems to be a stable convention concerning the question of existence, as well as a correct and non-arbitrary way to determine the convention of reference in the context. For example, "Catoblepas exist" is false because, as far as ontology goes, we've officially opted for eliminativism. Contrast this with "Memory exists" or "The table is solid," which both seem true in nearly all contexts, since (arguably) we officially take a revisionist approach to memory and solidity.

The *origin* of a convention can be rather arbitrary, as when [astronomers vote](#) on how to use the term "planet." But normally an individual can't simply shift the convention in an instance based on convenience. (Sorry, [Humpty Dumpty](#).) Thus, it doesn't seem we are licensed to adopt an eliminativist approach to an entity whenever it suits us, even if reference conventions can change with context. Rather, the reference convention is stable for certain kinds of questions (e.g. "Do catoblepas exist?" versus "Are catoblepas just wildebeests?").



There are two problems here. *First*, the semantics of the term "free will" that Nichols ends up with doesn't seem to be analogous to the models on which it's based. *Second*, even if

we could find a model that involves arbitrariness in the choice of reference conventions (e.g. descriptive vs. causal-historical), it would seem an implausible model for freedom and responsibility.

At any rate, I wonder if other theories can avoid either problem. For example, a classical [contextualist](#) view might be flexible enough to provide a model for how the reference convention can vary across many contexts. Of course, that would require a story about which standing linguistic meaning governs reference in different contexts. Yet even [indexicals](#), which are quite shifty, have a standing meaning once the convention is fixed. For example, the linguistic meaning of "I" is something like "the speaker," the reference of which shifts with context. But what could the linguistic meaning of "free will" be such that the type of reference convention shifts *arbitrarily*?

Alternatively, we might treat the concept of free will as a cluster concept, modeled say on either [prototype](#) or exemplar theory. Indeterministic choice might be one component among many in the ordinary notion of free will, but it isn't a necessary condition. (Full disclosure: [I've defended this approach](#) to the ordinary concept of free will, [as mentioned on this blog](#).) Such cluster theories are psychological, not semantic, so they can capture stable and non-arbitrary reference conventions for certain types of statements involving the relevant term. Moreover, a cluster theory avoids attributing a mere verbal dispute between compatibilists and incompatibilists; it treats both positions as being to some degree intuitive; and yet it allows that one side's intuitive pull can be debunked.

I'd love to hear what others think---not just about my worry but also about Shaun's fabulous book! And I think people should feel free to comment on the above, even if they haven't read the book.