EuJAP | Vol. 15 | No. 2 | 2019
UDK: XXX
https://doi.org/10.31820/ejap.15.2.5

HOW DO WE KNOW THAT WE ARE FREE?

Timothy O'Connor
Indiana University

Review article – Received: 14/06/2019 Accepted: 02/11/2019

ABSTRACT

We are naturally disposed to believe of ourselves and others that we are free: that what we do is often and to a considerable extent 'up to us' via the exercise of a power of choice to do or to refrain from doing one or more alternatives of which we are aware. In this article, I probe the source and epistemic justification of our 'freedom belief'. I propose an account that (unlike most) does not lean heavily on our first-personal experience of choice and action, and instead regards freedom belief as a priori justified. I will then consider possible replies available to incompatibilists to the contention made by some compatibilists that the 'privileged' epistemic status of freedom belief (which my account endorses) supports a minimalist, and therefore compatibilist view of the nature of freedom itself.

Keywords*: Free will, freedom experience, incompatibilism, a priori justification, conscious awareness, revisionism*

1. Introduction

We human beings are naturally disposed to believe of ourselves and others that we are free: that what we do is often and to a considerable extent 'up to us' via the exercise of a power of choice to do or to refrain from doing one or more alternatives of which we are aware. In what follows, I will probe the source and epistemic justification of our 'freedom belief'. I propose an account that (unlike most) does not lean heavily on our first-

personal experience of choice and action, and instead regards freedom belief as a priori justified. I will then consider possible replies available to incompatibilists to the contention made by some compatibilists that the 'privileged' epistemic status of freedom belief (which my account endorses) supports a minimalist, and therefore compatibilist view of the nature of freedom itself.

2. The Source and Justification of Our *Freedom* Belief

I start from the large but widely-shared assumption that our belief in agential freedom ('free will') in mature human beings is somehow or other 'properly basic', rationally warranted independent of any evidential connection to other warranted beliefs.[1] My aim is merely to determine the most plausible account of *how* this is so.

A common view among philosophers past and present is that our belief in freedom is based in an *experience as of freedom* that pervades deliberate choice and action.[2] If this is correct, we may readily propose an analogy with beliefs that have their immediate source in sensory experience. It is widely held that, e.g., my sensory-based belief that I am sitting in a chair is non-inferentially rationally warranted, despite both its being conceivable that I am dreaming and the fact that my perceiving a chair *as* a chair

[1] A philosopher of a strongly empiricist bent might propose instead that our freedom belief is rooted in third-personal evidence of systematic connections between our antecedent psychological states, our choices, and our subsequent actions. But I doubt that such evidence is robust and specific enough for this purpose unless one endorses a deflationary view of the content of our freedom belief.
In a variation on such an account, Nichols (2015, 42-49) suggests that each of us makes a statistical/inductive (or possibly deductive) inference from our *own* case in coming to think that our choices are causally undetermined (he does not distinguish, as I do, freedom belief from the belief that choices are causally undetermined). Our not being *aware of* determining causes of our decisions (in typical cases) is paired with an assumption that all causal influences on decisions are introspectively available, yielding the conclusion that they are not determined. But again it seems to me psychologically implausible that we each come to form and sustain such belief on broadly empirical grounds. Nichols acknowledges that there is no direct evidence that this is so. Instead, he claims that it provides a 'how possible story' that in the absence of any other good explanation is a plausible contender for being the correct story. I go on to give a different account that better meshes with the fact that freedom belief is widespread, if not universal, and is implicated in our moral outlook. Our practice of moral accountability is plausibly more deeply rooted in human psychology than this kind of inferential story would indicate. Nichols further claims that, if his hypothesis concerning the inferential origin of belief in indeterminism is correct, its rational warrant is undercut by scientific evidence of unconscious causal influences. On the evidential bearing of unconscious influences on belief in indeterminism that is *not* inferred in the manner Nichols proposes, see fn.12 below.
[2] For a recent defense of this view, see Guillon (2014, 2017). See also Holton (2009).

depends causally on my having had prior experiences and conceptual learning. Beliefs stemming directly from sensory experience (or many of them) are epistemically 'innocent until proven guilty'. Likewise, it may be claimed, for our belief in our own freedom, grounded in an experience as of freedom.

To assess this proposal, we need to consider the content of our experiences as of freedom. Psychologists have suggested that the background/focal distinction that is apt for describing sensory awareness also applies to awareness of our own agency (Wegner 2002). When I walk to campus along the usual route, I am often thinking about the lecture I am about to give. I barely attend to my stopping at the traffic light or my continuous action when not so stopped of moving my legs. Nonetheless, I have a background sense of being in control of what I am doing.

It is difficult to characterize precisely this background sense of agency, though we'll return to it below. What most philosophers have in mind when appealing to experience as grounding warranted freedom belief is not this background sense of agency but instead a more focal and episodic experience: the experience we have when consciously and more or less deliberately deciding what we shall do when confronted with a limited number of action alternatives. In such cases, it seems to me that it is in my power to determine the choice I am about to make – at a minimum, a power to do or not to do some contemplated action.

It is not sufficiently appreciated that an experience-based account of the epistemic warrant of freedom belief must make several tacit empirical commitments.[3] The most obvious of these is that the experience as of freedom is a cross-cultural universal, rather than being limited to those who have been reared in particular cultural ways of thinking about agency and responsibility. There is some evidence in support of the universality of freedom experience (Sarkissian et al. 2010), but it remains to be firmly established.

A second empirical commitment is that such experience, even if universal, is the basis of freedom belief, rather than the other way around, and that it is also not substantially shaped by any other explanatorily-prior belief, such as a belief in moral responsibility. Against this, one might point to evidence that the degree of control one self-ascribes can be modulated to some degree by external cues (Desantis et al. 2011, cited in Bayne 2016, and Wegner 2002). However, such studies are limited (for feasibility reasons) to post-choice reports, rather than targeting real time

[3] See related discussion in Bayne (2016, 641-642).

experiencing-in-willing, and so provide direct evidence only for the malleability of post hoc beliefs.

A third assumption that this epistemic account appears to require is that the experience as of freedom is appropriately causally related to the process of choice and action. Sensory experience is a reliable causal consequence of the physical reality perceived; likewise, it seems, freedom experience, if it is to ground the justification of freedom belief, should be reliably and fairly directly caused by (if it is not simply an aspect *of*) the formation of choice – the manifestation of the power seemingly experienced. Some see evidence to the contrary in certain abnormal clinical phenomena such as anarchic and alien hand syndromes, in which an individual engages in purposive behavior (e.g., reaching for someone else's glass of water) while lacking the experience as of controlling (or even desiring) the behavior. The conclusion drawn is that the causal pathway of the experience as of freedom is quite distinct from the origin of purposive decision, and so such experience (when present) should not be taken to be a plausible epistemic basis for justified belief concerning the nature of purposive action itself. Note, however, that this establishes only that *purposive action* can occur without freedom experience, and we already knew *that*. Purposive action is a broader category than directly free action, encompassing the significant portion of our behavior that is automated, including the routine behavior noted above of taking a familiar route to work each day.[4] Usually, such behavior is also accompanied by a background sense of agency, and that is what is missing in these clinical cases (to the agents' considerable distress). But neither the diffuse background sense of agency nor the unconsciously generated and regulated behavior it normally accompanies are at issue here. The theoretical commitment of the epistemic view we are exploring is that *deliberate conscious choices* very reliably either cause or have as a component an experience as of freedom in so choosing. The unusual cases cited simply do not speak to this claim. And even if cases could be adduced that prise apart these elements, unless there was reason to suppose that they do so with some frequency, they would not provide a compelling basis against the epistemological position that (as with the counterpart position regarding sensory experience) requires only substantial, not perfect reliability in the connection between experience and its object.

[4] Libertarians regularly make this distinction (see, e.g., Clarke 2003, 63, who distinguishes 'directly' and 'indirectly' free actions). Some compatibilists will dispute this, however, defining freedom of will and action in purely negative terms (the absence of certain freedom-undermining conditions). But such austere freedom theorists are unlikely for that very reason to give an experience-based account of the justification of our freedom belief, and so we may set their views aside for present purposes.

A fourth and final empirical commitment of an experience-based account of warranted freedom belief stems from the fact that each of us has first-personal experience only of our own agency and yet (unlike sensory experience) we would seem to have warranted belief in the freedom of others, too. This suggests the need for a two-part account on which belief in my own freedom is epistemically basic, while my belief in others' freedom is implicitly inferred from my belief that others are relevantly similar to me, including in their having experience as of freedom similar to my own. The latter clause commits one to a substantial empirical claim (about the source of a belief).

I do not see evidence that any of the four empirical assumptions has been significantly disconfirmed to date. But they are non-trivial assumptions that are much less evident than the corresponding assumptions we make regarding our own sensory experience. For this reason, it is desirable to have an account of the warrant of freedom belief that does not depend on these assumptions.

Such an alternative account is ready to hand: rather than drawing an analogy with belief rooted in sensory experience, we may draw one with our foundational empirical belief in a regular causal order to physical reality. This is a belief that we bring to our experience and exploration of that reality – that serves as an unargued starting point for our investigations of that reality. Our belief in freedom, we may plausibly contend, is a starting point in our approach to *social* reality (cf. Strawson 1962), one facet of the 'theory of mind' that we are naturally disposed to apply when we attain an appropriate stage of cognitive development. Whatever its evolutionary origin, we are primed to see ourselves and our fellows as agents with a substantial measure of freedom of choice, which partly grounds our moral responsibility. This belief need not be grounded in an experience of freedom to have a privileged epistemic status, and it seems psychologically implausible that the belief first forms in individuals through inference from freedom experience. That said, this is ultimately an empirical question; an account on which freedom belief occurs and is warranted *independently* of freedom experience incurs an empirical commitment no less than account on which there is a dependence. We'll be on safest grounds if we endorse the disjunction of the two, with the choice between them to be resolved (if it can be) on empirical grounds. One way to draw the two accounts closer together is to suppose that freedom experience is a significant part of the developmental *trigger* on a

disposition latent in our cognitive architecture towards freedom belief (in ourselves and others).[5]

3. Justified Freedom Belief and 'Risky' Theories of Freedom

Incompatibilists hold that the falsity of causal determinism is a necessary condition on our being free.[6] Some compatibilists contend that only a successful, final physical theory with the implication of causal indeterminism could give us reason to believe that indeterminism obtains. As the jury is still out on what a final physics will imply, we ought to be agnostic about whether our behavior is determined. But, they go on to argue, since we *are* entitled to believe that we are free, we have reason to think that compatibilism is true, since its truth, unlike that of libertarianism (the conjunction of incompatibilism and the thesis that we are free), is independent of this still-open question.[7] Put another way, libertarianism has implications for physics and neuroscience (the science most directly

[5] I thank Michael Murez for helpful discussion on this point. Jean-Baptiste Guillon pointed out to me that the account I am suggesting leaves an epistemic gap between 'people often act freely' and '*this* action was freely performed.' I am inclined to think that we close this gap in practice by noting the circumstances of the action, and in particular the experience of uncertainty as between alternatives. In making this suggestion, I am further amalgamating the two accounts.

[6] Parallel to disputes regarding the source of *freedom* belief, there is disagreement among libertarians regarding the source of their epistemic justification for believing that our choices are causally undetermined. Some say that this, too, is directly given in the experience of making deliberate choices. Most compatibilists will concede that it is *not* part of the content of our experience of making a deliberate choice that my choice is causally determined: there is no experience as of factors being causally sufficient for producing our choices. More controversial is the contention of some incompatibilists that we have the experience as of *not being* causally determined – that our agential experience has 'libertarian content.' The concept of causal determinism is of course too sophisticated a concept to plausibly attribute it to the explicit content of universal, mature human experience. A more plausible claim is that the best *articulation* of our somewhat inchoate experience as of freedom entails that, if it is veridical, our choices are not causally determined. It is the experience as of a 'two-way' (or multi-way) power to settle what our own motivations do not, and a satisfaction condition on the reality of such power is that our choices are not causally determined. I myself regard this claim as plausible, but it is controversial and difficult to adjudicate. Other libertarians would say instead that the belief that freedom requires causal indeterminism is justified solely through theoretical inference from, e.g., some version of the Consequence Argument. (Defenders of the former position might connect the two by maintaining that debate over the soundness of the Consequence Argument for incompatibilism as at root a dispute regarding the content of our own experience of freedom in action.) It is not my purpose to argue a position on this matter here.

[7] "One of the main virtues of compatibilism is that [its] most basic views about our agency—our freedom and moral responsibility—are not held hostage to views in physics" Fischer (2007, 81).

germane to the etiology of human action). But we have no business believing in advance of the science that the best final theories in these domains will have nondeterministic dynamics.

I will now consider three replies that libertarians have made to this argument and then propose and endorse a fourth.

1st response: compatibilism has scientifically risky commitments, too

Libertarian accounts of (direct) freedom differ, but they often have the form of endorsing many conditions commonly recognized by compatibilists and then adding *at minimum* a condition of significant causal indeterminism. Therefore, let us concede that compatibilist accounts of freedom require less than libertarian accounts. (In reality, this issue is slightly clouded by the fact that some compatibilist accounts impose conditions rejected by others. A given libertarian account may build upon one of the less stringent compatibilist accounts, and so not require a condition imposed by another compatibilist account, and not all compatibilist conditions are obviously met by all free human persons – Frankfurt's (1971) hierarchical account comes to mind.) The first response contends that it is not only distinctively incompatibilist conditions on freedom that seem potentially falsifiable by future science. In fact, recent studies in cognitive and social psychology have been claimed to show that human agents are badly ill-informed about their own motivations for acting as they do and, furthermore, that their experience as of consciously willing to act as they do is neither an aspect of nor caused by the actual, unconscious processes that generate their behavior.[8] Admittedly, the arguments made from such studies are overblown,[9] but (says the first respondent) the very fact that competent and knowledgeable theorists wish to debate these claims shows that they are not scientifically innocent. Libertarians may be "hostage to" views in future physics, but insofar as (many) compatibilists endorse conditions on freedom that these recent contentions have put on the menu for scientific study of human action, they are hostage to views in psychology.

However, I think the compatibilist has a reply here that is not available to the libertarian. For it is hard to see how science *could* consistently deny the *efficacy* of our conscious wills as a general matter. Scientific theories, models, and results are themselves the products of scientific *activity*: of human persons acting in certain coordinated, purposive ways and communicating their activities and results to one another. While the reality

[8] For an engaging, if slightly dated overview of many such studies, see Wegner (2002).
[9] See O'Connor (2009) and Mele (2009).

of reliably-known, purposive action may not be an explicit *premise*, or part of the theoretical *content*, of scientific theories, it is a *pragmatic assumption* of such science: if we supposed it to be false, we would thereby have reason to doubt the trustworthiness of the outputs of such activity. It is reasonable to accept the trustworthiness of these outputs only insofar as we take them to have resulted from actions guided by the specific conscious purposes and beliefs that the actors' report them to have been. To *deny* the efficacy of conscious will is to saw off the branch on which one sits. One certainly may argue unproblematically that human action and self-awareness are prone to error and ignorance in a variety of specific forms. Our grasp of our own motivations is imperfect, we are sometimes self-deceived, and it is not always easy to come to a more accurate self-understanding even when we learn of the flaws in our cognitive design. 'Willusionism',[10] by contrast, is inherently unstable because of its sweeping generality, as it thereby encompasses the very activity of the would-be unmaskers of human agency. (This simple point is not sufficiently appreciated by some 'no free will' scientists who precisely target at times the assumption of conscious efficacious agency, which they do not clearly distinguish from freedom as libertarians understand it.) This is, if you like, a transcendental argument for effective conscious agency, but not for libertarian freedom.

A libertarian might contend that scientific practice presupposes indeterminism also, in the form of real alternatives open to the scientific investigator in experimentally probing and manipulating natural processes. Scientific experimental interventions are deliberate attempts to impose a *departure* from the natural, law-governed unfolding of events, suppressing some natural dispositions and artificially stimulating others in an effort to isolate and characterize causal variables not previously understood. But it is far from clear that this conception of experimental interventions entails a departure from fundamental, deterministic regularities. They may, rather, belong to a special kind of macroscopic process that is determined to occur in accordance with psychophysical law – one part of Nature causally determined to query the whole, not producing events that depart from what Nature as a whole was bound to do, but rather events that depart from the kind that would have occurred in the absence of such intervening systems (and where such absence then and there was itself precluded by prior events).

[10] An apt term coined by Eddy Nahmias (2011) for the view that the experience of efficacious conscious willing is a pervasive illusion.

2nd response: the limits of conclusive confirmation of deterministic theories

The first response to the compatibilist's challenge that the libertarian's fortunes are implausibly hostage to future physics was to contend that a similar challenge is faced by most varieties of compatibilism. A second response is to argue that neither view faces such a challenge, as it is a paper tiger. There is no threat because there are inherent limits to what science can establish when it comes to anything as complex as human agency. Dynamical theories about elementary phenomena (such as quantum mechanics) draw most of their evidence from studying the behavior of small systems in artificially isolated contexts, near vacua where external influence is screened off. But libertarians do not (typically) accept the reductionist premise that human beings and their behavior are simply the resultant of trillions of micro-interactions of their simplest parts and those of their surrounding environment. They (and some compatibilists) will suppose that freely made choices in particular are strongly emergent phenomena, where this entails a kind of 'top down' control of certain highly organized systems over their own behavior. This strong emergentist thesis is not disconfirmed by the successes of particle physics in accurately and fully describing the behavior of matter in simple, non-organized contexts.

This reply is, I believe, cogent as far as it goes: the question of whether human choice is fully causally determined will not be settled by the character of an ex hypothesi 'final' physical theory. However, there is a better candidate science for (eventually) giving significant evidence in favor of the determinist option on the question, and that is neuroscience, assisted by more functionalist branches of cognitive psychology. The challenges it faces in the attempt to settle this question are not trivial: there are 80-100 billion neurons in the mature human brain, with many hundreds of millions likely involved in regions directly impinging on human choice dynamics. There is the open question of indeterministic quantum effects bubbling up from below to be grappled with, as well as getting a theoretical grip on what plausible and detailed emergentist hypotheses might look like. And, independent of these complications, we are a long ways off from having any kind of testable and detailed theoretical hypothesis concerning the neural process underlying human deliberation and choice, which may well be subject to significant individual variability. All that acknowledged, one can imagine a feasible development of the science to the point that regions of the brain of a deliberating person could be monitored in real time with *sufficient* fineness of grain to yield psychological correlates of measurable strength that enable testable predictions of behavior in

paradigm ostensible instance of free choice.[11] Doubtless such studies would require approximating techniques and less-than-certain assumptions that could be disputed. But no one has offered a compelling reason to think that it will be infeasible indefinitely for the science to advance to the point where significant *evidence* that human deliberation approximates a deterministic process might be adduced.[12] I conclude that our second possible response, too, is unsatisfactory.[13]

3rd response: hedging one's bets on incompatibilism

Peter van Inwagen (1983, 219-221) reports that his various *a priori* commitments in the matter of free will and moral responsibility are of variable strength. In particular, his confidence in *we are morally responsible creatures* is greater than it is in *we have free will* which is greater in turn than it is in *incompatibilism is true* and *some of our acts are causally undetermined*. This leads him to suggest that, if determinism were empirically established, he would abandon his incompatibilism, leaving intact his other, stronger commitments. In reply to the compatibilist charge that his incompatibilism renders his beliefs concerning moral responsibility and freedom "hostage to" physics, he in effect says that only his incompatibilism is so hostage, not his commitment to the reality of responsibility and freedom.

Let us consider van Inwagen's stance more carefully, in order to determine whether it is one that libertarians generally might plausibly endorse. Van Inwagen contends that his strength of belief in the following propositions are ordered (stronger to weaker) as numbered. (His belief in (3) and (3a),

[11] I leave aside discussion of Benjamin Libet's (1985) notorious conclusion from his famous study, since refined by many others right up the present day. The shortcomings of extant studies of this kind for addressing our present question have been made clear by many philosophers (e.g., Mele 2009), and recent scientific work has called into question precisely what kind of neural process Libet studies are tracking (beginning with Schurger et al., 2012).

[12] Terry Horgan (2015) and Tim Bayne (2016, 641-2) mistakenly claim that there is ample evidence against libertarianism *already*, in that cognitive science indicates myriad unconscious influences on human choice. But this is a very weak argument, since libertarians do not, as a rule, deny that we are subject to such causal influences. They are committed to denying only that such factors collectively *determine* all our choices.

[13] It is open to the proponent of the second reply to argue that our a priori justification for believing in the conjunction of incompatibilism and the belief that we are free is sufficiently strong that it would necessarily outweigh such an inference to the best explanation in favor of determinism based on somewhat indirect evidence. But even such a contention would need to concede that strong but defeasible evidence for determinism would require us to *weaken* our confidence in our belief in freedom. Further discussion of this general point occurs in my discussion of the third reply, immediately following in the text.

and (4) and (4a), are equally strong, with the 'a' propositions being a direct consequence of the similarly numbered propositions and one above it.[14]):

> (1) We are sometimes morally responsible for the consequences of our acts;
>
> (2) The validity of Beta entails that our having free will entails indeterminism;
>
> [Beta is the key 'transfer' of inability principle in his argument for incompatibilism. So van Inwagen is saying that Alpha and the other, 'fixity' premises are *more* certain than Beta, which comes in at (4).]
>
> (3) If (1) is true, then we have free will;
>
> ['Free will' for van Inwagen is having the ability to act other than what one does; this proposition is the Principle of Alternative Possibilities.]
>
> (3a) We have free will;
>
> (4) Beta is valid;[15]
>
> (4a) Our having free will entails indeterminism;
>
> [The thesis of Incompatibilism]
>
> (5) Indeterminism is true. (219)

Although he 'prefers' the propositions in this order, van Inwagen regards the conjunction of them as 'very likely' and so each of the conjuncts as very likely. He thus thinks it *very likely* that indeterminism is true in particular. But he goes on to say that if he were persuaded that science gave him an indisputable reason to accept determinism, he would reject Beta (4) and Incompatibilism (4a), since the (ex hypothesi) false (5) follows from (3a) and (4a), and he prefers (3a) to (4a), and (4a) itself follows from (2) and (4), and he prefers (2) to (4). So, the equally likely and linked (4) and (4a) would both have to go. He adds, crucially, "[a]nd that would seem to be the end of the matter" (221).

In conversation, some philosophers have expressed puzzlement at van Inwagen's conditional response to learning the truth of determinism, on the

[14] By this same reason, van Inwagen should have labeled (5) as "(4b)." I query this reason below.

[15] Van Inwagen has come to accept that Beta is invalid, but he now accepts a successor principle that functions much the same in the argument for incompatibilism.

grounds that the denial of (5) is a straightforwardly empirical claim, and that should not be the primary grounds for abandoning a purely conceptual claim such as (4), which is necessarily true, if true at all.[16] (1) and (3), as other empirical claims, are better candidates for being disconfirmed by the falsity of (5). But the general constraint on evidential support does not seem correct, as is shown by the following simple example[17]: I reason from purely mathematical principles, some uncontroversial and others less so, that Fermat's Last Theorem is false, and I am confident but less than maximally certain of my reasoning. Then my trustworthy friend Andrew the esteemed mathematician tells me that the theorem is true (and nothing more). It seems that I can reasonably be led on this empirical basis (simple testimony) to abandon the conjunction of the less-certain propositions.

A significant point of disanalogy is that in the mathematical case, my conclusion is derived from only putatively necessary premises, whereas in van Inwagen's case it is a mixture of an empirical claim and modal claims.

[16] Fischer (2016, 48) initially frames his 'problem of metaphysical flip-flopping' this way ("the rejection of an *a priori* ingredient in the incompatibilist's argument, contingent upon learning that causal determinism is true," 48), but he develops his criticism of van Inwagen's stance in different terms. His first considered criticism is that causal determinism is 'evidentially unrelated' to the crucial principle 4 (Beta), and so learning the former ought not to affect his commitment to the latter (54). This is uncompelling. Learning something may reveal to us that at least one of a small set of beliefs must be false, without making clear which. Fischer goes on to object to van Inwagen's preference *ordering* for the reality of moral responsibility over the principles that are needed to infer indeterminism. While I, too, find this ranking somewhat unnatural, it's hard to make a case that such a preference is irrational. Further below in the text, I note that the controversial status of the principles may well lead one to be less than maximally confident in them.) I go on to suggest that the real problem with van Inwagen's stance is his apparent commitment to the unrevisability of his belief in moral responsibility. Fischer expresses something similar in maintaining that van Inwagen should be open to the option of moral-responsibility *skepticism*, but that is different - and an odd complaint from one who endorses the objection to incompatibilism that set the stage for our consideration of van Inwagen's response! The way out that goes overlooked by van Inwagen and (here, at least) by Fischer is the option of being open to a form of revisionism when it comes to moral practice, which I develop near the end of the paper.

[17] I find van Inwagen's own reason for rejecting it unconvincing: "I have defended (Beta) entirely on a *priori* grounds. But it would not surprise me too much to find that this proposition, which at present seems to me to be a truth of reason, had been refuted by the progress of science. Such refutations have happened many times" (221). Presumably he is alluding to examples such as the rejection of Euclidean geometry by the Theory of General Relativity, or the Principle of Sufficient Reason (PSR) and Quantum Mechanics. A more accurate interpretation of this history, it seems to me, is that purely conceptual developments enabled thinkers to see possibilities hitherto unimagined (the separability of the particular parallel postulate from the other axioms of Euclidean geometry and their consistency with alternatives; the coherence of irreducibly statistical forms of explanation, allowing for a formally weaker but no less universal regulative explanatory principle than PSR), and this conceptual space was then exploited by empirical theorists. But nothing in the text hangs on my disagreement with van Inwagen on this point.

Might we suppose that in the latter kind of case, empirical evidence ought to lead to revision only of empirical claims in the former basis for the disconfirmed proposition? But doing so would seem to require setting aside van Inwagen's believing the empirical claim (we are morally responsible) more strongly than the putative truths of reason.

To take things further, let us consider another couple analogous cases:

> *BIV:* (1) This is a hand; (2) *this is a hand* entails *I am not a brain in a vat*; *so* (3) *I am not a brain in a vat*. I learn that (3) is false.
>
> *Martian:* (1) We are sometimes morally responsible for the consequences of our acts; (2) if (1), then our acts are not all a more-or-less direct product of remote Martian manipulation via secret micro-chip brain implants; so, (3) our acts are not all a more-or-less direct product of remote Martian manipulation via secret micro-chip brain implants. I learn that (3) is false.

Suppose that, for each of the cases, a philosopher believes proposition (1) more strongly than she believes proposition (2), although she judges each of them to be very likely true. And she further believes that were she to learn not-(3), she should reject (2) and retain (1). This would not be a mystifying stance – it could be held on the basis of a not-crazy theory about the role of reference in determining meaning – but I would regard it as implausible nonetheless.[18] In the imagined, extreme circumstances, it seems more reasonable for me to abandon (1) rather than the conditional expressing one of (1)'s evident implications. And so, I expect, would nearly everyone judge. (Van Inwagen himself uses the Martian example against the 'Paradigm Case' defense of compatibilism.) That indicates, though, that, with respect to each case, I believe (2) more strongly than (1). One question, then, is whether van Inwagen can reasonably hold a different preference ordering in the original case, believing in moral responsibility more strongly than he does in the conditionals expressing its putative theoretical implications (PAP, Beta and Incompatibilism). Note that in this case, there is nothing approaching universal agreement on those alleged implications, unlike (perhaps) the counterparts in *BIV* and *Martian*. Convinced but reflective incompatibilists such as van Inwagen might take this sociological difference to reflect a difference in 'closeness' of the theoretical commitments to the pre-theoretical concept of moral

[18] See Heller (1996) for just such a response to the Martian case. Deery (2019, msp. 11-13) shows how one can embrace a more nuanced causal-historical theory of reference for the concept of free action without concluding that we are free if the Martian control scenario were actually the case.

responsibility (and freedom). Further, as on most questions of degree, incompatibilists will differ in their precise judgments in these matters, with some seeing a tighter connection than others.

So far, we have not seen a convincing reason to regard van Inwagen's stance as an unreasonable one. However, even if van Inwagen reasonably assigns credences as he indicates, it does not follow that his method for handling evidence conflicting with a strongly held belief is correct. There are options beyond continuing to believe or coming to reject beliefs that underlie one's disconfirmed beliefs, so merely identifying and repudiating the least strongly held such belief(s) that enable one to avoid outright contradiction at minimal cost would not "seem to be the end of the matter." A more fine-grained response looks for probabilistic evidential connections. ~(5) may not entail ~(3) or ~(1), but perhaps one with van Inwagen's commitments should judge that (3) or (1), or both, are less *likely* on ~(5) than they are on current evidence (which does not include ~(5)). Remember that we are considering a credence set (van Inwagen's) that regards *all* of (1)-(5) as 'very likely.' (Van Inwagen is a fully convinced, not half-hearted, libertarian.) If he comes to believe in determinism, he cannot rationally continue to affirm the conjunction of (1)-(4). But since his preference for (1) over (2), (3), or (4) is slight, and scientific evidence for determinism does not speak directly to *any* of them, it seems that the most reasonable belief revision is to downgrade his credence in *all* of them to some extent: he knows that at least one of them must be false, but he has no *firm* basis for singling out a particular one of them. Perhaps his continuing to believe (1) (which he antecedently believed most strongly of the four) can survive this revision, but it will be a less strongly held belief.

There may be a reason that van Inwagen doesn't consider this seemingly judicious stance. Note that van Inwagen regards (3) and (3a) as equally likely, and similarly for (4) and (4a). He says that he so regards these pairs of propositions because (3a) follows directly from (1) and (3), and (4a) follows directly from (2) and (4). But a logical implication of a pair of propositions should not be treated as equally likely as either of the individual propositions *unless one regards the other of the pair as certain*. To put it in probabilistic terms, just to make the point salient, if one assigns (A) a probability of .9 and a wholly independent proposition (B) a probability of .8, and A & B entail a distinct proposition C, which one believes *solely* on the basis of A&B, then one should add the chances of A's being false and of B's being false, and so conclude that C should be

assigned a probability of .7.[19] Van Inwagen's reported strength of beliefs (given their bases) are coherent only if he assigns probability 1 (or something *very* nearly it) to propositions (1) and (2), the 'more likely' propositions in the deductions of (3a) and (4a). Perhaps, then, van Inwagen treats (1) (the proposition that we are morally responsible) as a *controlling* proposition, something that we should hang onto, come what may – at least for all non-fantastical scenarios, such as the Martian case. The trouble with this stance is that it comes at the price that we must completely *sever* our commitment to moral responsibility from our commitment to any substantial claims regarding its empirical implications. And this simply does not sit comfortably alongside incompatibilist commitments. (As we saw above in considering the first response, it does not sit easily even with many varieties of compatibilism, although their empirical 'exposure' is more limited.)

4th response: belief in free will and moral responsibility is <u>defeasibly</u> a priori justified

A better response, I believe, pushes back more firmly against a central premise underlying the compatibilist's challenge, which earlier I expressed thus: "we have no business believing in advance of the science that the best final theories in [physics and neuroscience] will have nondeterministic dynamics." We are rationally entitled to many assumptions concerning ourselves and the causal character of reality in advance of scientific confirmation, starting with the reliability of the senses and memory and the regularity of the world's fundamental causal order. Nor is it clearly inconceivable that some of these rational and necessary assumptions might be falsified by future rational investigation. It seems conceivable, e.g., that the deep regularities of our world suddenly cease to obtain, being replaced by a quite different set of regularities, such that we come to realize that the world is partitioned into distinct aeons, individuated by distinct natural laws. (Our bodies depend on biological regularities, so it is challenging to see how *we* might survive across the transitional juncture. But it remains conceivable in 2019 that our bodies are not essential to us.) Certain of our beliefs that are justified *a priori* thus seem to be empirically *defeasible*. If we categorize our belief in freedom and responsibility in this way, we need not adopt the stance of proscribing future deterministic psychological theories. Instead, we are simply betting against them, while letting the chips fall where they may.

[19] Where one's confidence in C is not solely a consequence of one's confidence in A and B (and, as in the example, C is not equivalent to the conjunction of A and B) then probabilistic coherence requires only that one assign C a value between 0.7 and 1.0. (I thank Tim McGrew for pointing out an error I made on this score in a previous draft.)

If the combination of confident belief with allowing for only the barest possibility of its falsity seems improperly prejudicial, inimical to unfettered inquiry, one should be mindful of the piecemeal advance of science, especially in so complex a domain as human psychology. It is hard if not impossible to say which open lines of inquiry in psychology and neuroscience (if any) have the potential to lead to eventual significant disconfirmation of an incompatibilist conception. Major pieces remain to be put into place in our understanding of human psychology before such a big picture question will come squarely into view of mature science. And even if some lines of inquiry seem friendlier to our moral self-conception than others, we may be further mindful of William James' point more than a century ago that science is often helped, not hindered, by scientists having passionate commitment to competing perspectives that they seek to vindicate through rival research programs.

What, then, should we say concerning the hypothetical future scenario in which we come to believe that human behavior generally is, after all, psychologically determined? That the proper response would be to say, 'I guess we were wrong about all that' and to abandon moral practice altogether? I think not. This austere disavowal is not the sole alternative to van Inwagen's willingness to abandon his incompatibilism. There is a more attractive and fully reasonable stance for an incompatibilist that is in the spirit of van Inwagen's tenacity of commitment to moral responsibility. It is something like Manuel Vargas's (2007; 2013; see also Nichols 2015) *revisionism* – here taken as a hypothetical response to being given compelling evidence for determinism, rather than (as with Vargas) a current position. What precise shape a revisionist stance might take is a complicated question, one that needn't be adjudicated here to motivate the general stance. The basic idea is that, given evidence that our previous moral conception of human agency is unlikely or untenable while recognizing the centrality of moral thought and action to our practical lives, we might come to think differently (whether by choice or not) about what our commitment to freedom and moral responsibility should amount to, until a changed perspective begins to take hold and wholly supplants the previous way of thinking. There are our current associated concepts of freedom and moral responsibility, with their substantial empirical commitments, and there is a more general (and seemingly ineliminable) role that moral discourse plays in our practice. If push came to shove, that latter role could continue to be filled by retreating to the use of more modest, revised concepts that result from eliminating untenable elements

of the original concepts.[20] I do not say that the process of embracing such a revision would be a smooth one. Indeed, I think it would be deeply disconcerting to come to think that we are not free and responsible as we now understand those terms. But adjustment is merely difficult, whereas abandonment of practice seems psychologically impossible. Being disposed to go revisionist in the face of possible future empirical evidence against our current freedom and responsibility beliefs would allow one to agree with van Inwagen on the incompatibilist implications of our ordinary concepts, and to agree with him and many compatibilists on the practical 'unthinkability' of abandoning the practice of judging ourselves to exercise freedom in many of our actions and holding one another morally responsible for the consequences of such acts (in *some* recognizable sense), while also and more reasonably allowing that beliefs that have substantial empirical commitments should be disconfirmable. And once we recognize the availability and attractiveness of this more nuanced attitude regarding worst-case scenarios, we can fully meet the compatibilist's challenge.

I have proposed that our belief in our own freedom is epistemically warranted *a priori* while being defeasible. Whether it is grounded in regular experience as of acting freely is an open empirical question, but I am inclined to doubt it. (The thought that it *needs* to be so grounded in order to be rationally warranted is an empiricist prejudice that should be resisted.) I close by briefly responding to a skeptical query: if belief in our own freedom is instinctive and warranted *a priori*, whence occasional disbelief in free will among the intelligentsia? The natural answer is that this is a species of theoretical skeptical doubt, similar to skeptical doubts regarding, e.g., the reality of causation, another proposition that we are warranted *a priori* in accepting. In both cases, the theoretical doubt is matched by practical commitment to the thesis, expressed in behavior. This may involve the person's having contradictory beliefs. But another

[20] This of course assumes that not all elements of our freedom and responsibility concepts are essential to them. Fortunately, we need not resolve that question here. If this assumption is false, the revisionist proposal may take the form of *replacing* the original concepts with successor concepts that overlap the originals and that can still fill the broad role in moral practice that we cannot imagine abandoning altogether. For a map to possible forms that revision or replacement might take, see Nichols (2015, 59-62).
Deery (2019) proposes, alternatively, that *free action* is a natural kind concept and that we follow Boyd's (1999) analysis of such concepts as homeostatic property clusters, where not all properties in the cluster are essential to them, and where the applicability of the concept is consistent with our making significant false presuppositions concerning it. *If* it is widely and wrongly assumed that the properties we track with our freedom concept involve or require causal indeterminism (something Deery does not commit himself to), it would still refer. I doubt that this is the correct way to think about our freedom concept, and doubt more strongly that indeterminism is merely an implicit associated assumption concerning actions falling under the concept. However, the proposal merits further attention than I can give here.

possibility, and one that I find attractive, is that the person believes the target proposition while merely believing that he disbelieves (or fails to believe) it. That is, the theoretical doubt takes the form of a (mistaken) belief concerning one of the person's own first-order beliefs.

Either way, an advantage of the alternative, conditional revisionism suggested in the previous paragraph is that it would allow for continued coherence of one's practical and theoretical commitments.

REFERENCES

Bayne, T. 2016. Free Will and the Phenomenology of Agency. In *The Routledge Companion to Free Will*, eds. Timpe, Griffith, and Levy, 633-644. Abingdon, UK: Routledge.

Boyd, R. 1999. Homeostasis, Species, and Higher Taxa. In *Species: New Interdisciplinary Essays*, ed. R. A. Wilson, 141–185. Cambridge, MA: The MIT Press.

Clarke, R. 2003. *Libertarian Accounts of Free Will*. New York: Oxford University Press.

Deery, O. 2019. Free action as a natural kind. *Synthese*. https://doi.org/10.1007/s11229-018-02068-7.

Desantis, A., Roussel, C., and Waszak, F. 2011. On the influence of causal beliefs on the feeling of agency. *Consciousness and Cognition* 20, 1211-1220.

Fischer, J. M. 2007. Compatibilism. In *Four Views on Free Will*, eds. J. M. Fischer, R. Kane, D. Pereboom, and M. Vargas, 44-84. Oxford: Blackwell Publishing.

Fischer, J. M. 2016. Libertarianism and the Problem of Flip-flopping. In *Free Will and Theism*, eds. K. Timpe and D. Speak. Oxford: Oxford University Press.

Frankfurt, H. 1971. Freedom of the will and the concept of a person. *The Journal of Philosophy* 68: 5-20.

Guillon, J.-B. 2014. Van Inwagen on introspected freedom. *Philosophical Studies* 168: 645-663.

Guillon, J.-B. 2017. Épistémologie de la Causalité Agentive. In *Le libre arbitre: Perspectives contemporaines* [online]. Paris : Collège de France. <http://books.openedition.org/cdf/4939>. DOI: 10.4000/books.cdf.4939.

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

Holton, R. 2009. Determinism, self-efficacy, and the phenomenology of free will. *Inquiry* 52: 412-428.

Horgan, T. 2015. Injecting the Phenomenology of Agency into the Free Will Debate. In *Oxford Studies in Agency and Responsibility*, eds. D. Shoemaker and N. Tognazzini, 3: 34-61.

Libet, B. 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *The Behavioral and Brain Sciences* 8: 529-566.

Mele, A. 2009. *Effective Intentions*. New York: Oxford University Press.

Nahmias, E. 2011. Why '*Willusionism*' leads to 'Bad Results': Comments on Baumeister, Crescioni, and Alquist. *Neuroethics* 4: 17-24.

Nichols, S. 2015. *Bound: Essays on Free Will and Responsibility*. Oxford: Oxford University Press.

O'Connor, T. 2009. Conscious Willing and the Emerging Sciences of Brain and Behavior. In *Downward Causation and The Neurobiology of Free Will*, eds. F. R. G. Ellis, N. Murphy, and T. O'Connor, 173-186. New York: Springer Publications.

Sarkissian, H., Chatterjee, A., de Brigard, F., Knobe, J., Nichols, S, and Sirker, S. 2010. Is belief in free will a cultural universal?" *Mind and Language* 25: 346-358.

Schurger, A., Sitt, J. D., and Dehaene, S. 2012. An accumulator model for spontaneous neural activity prior to self-initiated movement. *PNAS* 109: E2904-E2913.

Strawson, P. 1962. Freedom and Resentment. *Proceedings of the British Academy* 48: 1-25.

van Inwagen, P. 1983. *An Essay on Free Will*. Oxford: Clarendon Press.

Vargas, M. 2007. Revisionism. In *Four Views on Free Will*, eds. J. M. Fischer, R. Kane, D. Pereboom, and M. Vargas, 126-165. Oxford: Blackwell Publishing.

Vargas, M. 2013. *Building Better Beings.* New York: Oxford University Press.

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