EUROPEAN JOURNAL OF
ANALYTIC PHILOSOPHY

UDK 101
ISSN (Print) 1845-8475
ISSN (Online) 1849-0514
https://doi.org/10.31820/ejap

Publisher, Editorial office
University of Rijeka, Faculty of Humanities and Social Sciences, Department of Philosophy
Address: Sveučilišna avenija 4, 51000 Rijeka, Croatia
Phone: +385 51 265 794
Fax: +385 51 265 799
E-mail: eujap@ffri.hr

Web address
https://www.ffri.hr/phil/casopis/index.html

Printed by Impress, Opatija (Croatia)
200 copies available

TABLE OF CONTENTS

SPECIAL ISSUE:
FREE WILL AND EPISTEMOLOGY

Special issue of EuJAP: Free Will and Epistemology

Guest editors:

László Bernáth
Hungarian Academy of Sciences and Eötvös Loránd University

András Szigeti
Linköping University

Timothy O'Connor
Indiana University

Preface to this Special Issue on Free Will and Epistemology by Robert Lockie (University of West London)

Let me begin by recording my gratitude to the editors of, and contributors to, this special issue. This volume follows on from two wonderful conferences on my 2018 monograph *Free Will and Epistemology* – one held in Budapest, one in London[1]. I wish to express my appreciation to all who contributed to these events, and to the institutions which hosted them – in particular, albeit with a heavy heart, the Central European University: subsequently all but driven out of Budapest by the calculatingly malicious actions of the present government of Hungary. As regards individuals, two of the editors of this volume (András Szigeti and László Bernáth) organised the former conference, whilst Tim O'Connor's paper grew out of his contribution to the latter. My deep and sincere appreciation to all.

Because of the compressed deadline for this preface, I have, at the point of writing, only had access to the abstracts of these papers, and, beyond recording my gratitude to the authors, must therefore largely refrain from introducing them further – I shall be reading them with great care subsequently and hope to respond to them in later work. Some of the papers

[1] My thanks to the Central European University (Budapest); The Hungarian Academy of Sciences (MTA); the Lund Gothenburg Responsibility Project; the Institute of Philosophy at the School of Advanced Study (University of London); the Mind Association; the Aristotelian Society; the University of West London; and especially to András Szigeti, László Bernáth and Tim O'Connor.

found herein are more addressed to the specific arguments found in my monograph, while others are more concerned with the titular issues shared by that book and this special issue. I wish here to situate these papers within the framework of this topic area as a whole – to motivate this area (free will and epistemology) as one of great and enduring importance to philosophy – and, if I may be permitted, of flagging a few of my own contributions in doing so.

Normative Epistemology and Free Will

There are numbers of recent works in the areas of free will or epistemology considered separately – both are of course currently flourishing research areas. Piecemeal connections between these two areas are widely acknowledged, and debates surrounding these issues are widely joined with regard to a number of distinct topics (e.g. engagement with the 'doxastic voluntarism' debates, the 'epistemic deontologism' debates, the 'reasons-responsiveness' debates, the 'does reflection presuppose open choice' debates). However, relatively little recent work exists which is at once an uncompromising contribution to both fields – work that is squarely situated within both sub-disciplines, as opposed to being situated in one sub-discipline and borrowing from, or making excursus to, the other. In particular very little book-length work exists which does this. I have argued that a historically and currently important position in normative epistemology (deontic internalism) has critical conceptual connections with an important position in the free will debates (libertarianism). Bluntly: that to be *epistemically justified* one must have *freedom of thought* – where this latter involves a strong notion of freedom, and the former involves a normative authority that is essential for reflexive epistemic justification. The work therefore requires participants in the free will / responsibility debates to effect serious engagement with epistemology – and vice versa. I am grateful to the editors of, and contributors to, this special edition for doing just that.

The Transcendental Arguments

One of the great metaphilosophical traditions is that of transcendental argument (*peritrope*, 'self-undermining' argument). Like all metaphilosophical traditions, this one is controversial. My book defends two, connected, transcendental arguments: one for a deontic conception of epistemic internalism (Part One) and the other for a strong notion of free will (Part Two) – with the latter argument relying in part upon the former. The latter is one of the great, famous, philosophical arguments – from Epicurus to

Kant to Popper. The second part of the book is an extended defence of this argument. In this special issue both Nadine Elzein and Toumas Pernu's co-authored paper and Simon-Pierre Chevarie-Cossette's paper assess instances of these transcendental arguments for free will – including my version – while Amit Pundik's paper argues that the transcendentally established notion of free will is so strong metaphysically that it implies the unpredictability of free actions.

The book argues that many determinist and some indeterminist accounts of free will are indefensible on the ground that they must withhold from their proponents the reflexive epistemic justification that these accounts themselves require. It likewise argues that certain epistemic views (radically externalist views – views constituting a 'totalising' externalism) are indefensible on the ground that they withhold from their proponents the reflexive epistemic justification needed to maintain these epistemic positions. The book recommends from this that we develop accounts in these areas that are reflexively defensible, and advances an account of epistemic justification ('thin deontological internalism') and an account of free will (self-determinism) which are just that. Relatedly, in this contribution, Luca Zanetti's paper investigates in detail whether this transcendental argument against externalism is successful after all. Andrew James Lantham and Timothy O'Connor map other novel ways to establish epistemic justification for believing in free will. Lantham argues that the careful analysis of the concept of free will will do the work; while O'Connor claims that one should consider the belief in free will as a belief that is justified a priori.

The "Thin Deontological" Account of Epistemic Justification

Part One of my book defends a currently rather unfashionable account of epistemic justification, one which was of extraordinary historical importance but has now partly fallen into desuetude. This strongly deontic notion of internalism was engaged with by Plantinga, Foley and Alston; and, going back further than these figures, has its roots in Clifford, Descartes, Locke, and much of the early-modern epistemological enlightenment. In this work, this is baptised as a 'thin deontological' notion of internalism, though Alston (1985) from the standpoint of a (guarded, partial) opponent, abbreviated this notion as 'J_{di}' – which stands for *deontic, internalist, justification*. Plantinga just calls this same notion 'internalism', but when pushed, *classical deontological internalism* – and deprecates those pure accessibilist internalists who depart from what he (an externalist) nevertheless identifies as its "deep integrity" (Plantinga 1993, 28). A version of this conception of epistemic justification has

become known as 'Foley Rationality' (cf. e.g. Foley 1993), while Bergmann (2006) entitles it 'subjective deontological justification' or 'epistemic blamelessness'. Other major figures (Chisholm, BonJour, Goldman) played major roles engaging with this notion throughout the 1980's. Although epistemic deontology per se is currently quite well represented in recent epistemology, its defenders tend to be insufficiently rooted in the ethical literature, and tend to fail to follow through the 'ought' implies 'can' *ethics of belief* entailments of said view to their logical, perspectival, conclusions. They also tend to be insufficiently 'metaphilosophical' in their epistemological purview – and thereby insufficiently reflexive in their practice. It is regrettable that this highly motivated and carefully thought-through variant of epistemic deontology per se, with its deep historical provenance and elegant connections with the ethical literature, appears to have been substantially marginalised or eclipsed; and (apart from Foley's ongoing work) has rather wanted for recent defences. I defend and deploy this subjective, perspectival, deontically internalist notion of epistemic justification – adverting as it does to a deep, neo-Cartesian 'ethics of belief' tradition – whereby justification is taken to involve the discharge of one's *epistemic responsibilities*, as *dutiful thought*, as *reasoning as one ought*.

Resisting Transcendental Arguments

One of the first points to make – or rather concede – is that, at a superficial level, it is very easy to resist transcendental arguments of the kind I advance in my work. For instance, if one wishes to claim (as I do) that we cannot be justified in abandoning deontic epistemic internalism because the 'last ought' is the ought which urges us to abandon all oughts – or, more prosaically, that one cannot abandon an *oughts*-based epistemology *tout court*, since one would have to claim that one *ought* thus to abandon said epistemology – the obvious response would be to contend that this is question-begging. One merely embraces an alternative, non-oughts-based notion of justification and uses this to effect the abandonment. Were the counter really that obvious, why advance such a transcendental argument at all?

However, in the face of such 'question-begging' objections, a number of issues arise. One question is whether there is such a notion as that to which these objections make appeal: that is, a notion genuinely of *justification* ('our concept', justification itself, the *Echt* notion thereof, and not some other, more-or-less *Ersatz*, more-or-less revisionary thing). Is this justificatory notion *radically* (wholly, at every level, without remainder or

concealed indebtedness) non-deontic?[2] May we take one such *purely* non-deontic yet *genuinely* justificatory notion 'off the shelf', as it were? Has the proponent of this 'question-begging' counter appreciated deeply enough that it is *wholesale* ('totalising') replacement that needs defending here? Of course, there are non-deontic notions in epistemology – I defend and employ such notions in my book. Of course, they are of great importance in epistemology. But they function in an epistemology in which they are seen as not the *only* normative kinds. The question is whether we can be reflexively justified given the wholesale, *totalising* abandonment of any notion of epistemic 'ought' – or, put another way, whether we can avail ourselves of these other notions (of truth, reliability, access, mentalism, 'objective' rationality, etc.) to do *all* the work our former notions did, without *at any point* needing to make appeal to *reasoning as we ought*. That is, (as for the case for *Meno* road-to-Larissa cases of truth simpliciter) without merely ceasing to do epistemology – or at least, the epistemology of epistemology, epistemology where this concerns terminus issues of justification and not some other thing. Where do we repair to if we thus abandon said (deontic) notion of justification *tout court*? That is, *reflexively*, at every level, how do we effect the tasks which formerly were effected by this deontic notion – now, supposedly, to be replaced?

As an example of how dismissive such 'question-begging' counters can be, consider the attempt to respond to a transcendental argument in a different area (eliminativism) by Paul Churchland (I responded to, and quoted this passage, in Lockie 2003). Having urged that we abandon no less than *beliefs, desires, consciousness, truth, reference, rationality, sentences, logic, language*, Churchland responds to *peritrope* objections (e.g. from Lynne Rudder Baker, that this would constitute an act of 'cognitive suicide') in the following terms:

> Let us concede then, or even insist, that current [folk psychology] permits no tension-free denial of itself within its own theoretical vocabulary. [...] [A] *new* psychological framework [...] need have no such limitation [...] we need only construct it, and move in. We can then express criticisms [...] that are entirely free of internal conflicts. This was the aim of [eliminativism] in the first place. (Churchland 1993, 214).

"We need only construct it, and move in" – well, that's rather breezy is it not? Did we really "construct" our previous framework? And is our new

[2] Or is it, rather, putatively 'deontic' yet seen as devoid of 'ought' implies 'can' entailments which had hitherto appeared *internal to the concept* of deontology itself and as such?

home available to order as a development, "off plan", as it were? Can we live in these (envisioned, advertised) new premises – are they habitable, for creatures like us? What are their specifications? Will our new dwelling offer us all the qualities of our previous living space? Presumably not – that was the point of the replacement, was it not? But then, is there an overarching perspective affording a view which permits of normative comparison between the properties of our former (actual, extant) normative framework and some envisaged, not-yet-in-existence *Philosophie der Zukunft*? Problems both of incommensurability and an inherent tension between "view from nowhere" and "there is no view from nowhere" commitments ineluctably bedevil any such philosophy of wholesale vast normative replacement: problems that do not suggest such a blasé response would be by any means easy to defend.

Eliminations of great normative frameworks require a great deal more than simply pointing to an existing framework and saying, as it were, from outside of it, or sideways on to it: "I have decided to abandon *that*!" Serious philosophical and metaphilosophical work is needed to establish whether any such abandonment is possible, or even conceivable, much less feasible; and what would follow were this so – which of our practices could be preserved, which would need revision (and to what extent) and which would have to be abandoned. Serious work would also be needed to consider the knock-on, holistic chasing-through of the unintended and unforeseen consequences of said revisions and abandonments. The avowed presence of existing *piecemeal* alternatives to a given normative philosophical account (e.g. epistemic externalism vis à vis deontic internalism) *given that one is not seeking a wholesale, 'totalising' elimination, with inevitable commitments at the reflexive, metaphilosophical level*, does not establish the viability of said wholesale, totalising, elimination. This is a point well appreciated in the Lucretian continuation of the Epicurean tradition of *peritrope* argumentation within epistemology: one may claim one can doubt any one thing without thereby establishing one can doubt everything (Lucretius 1947, Bk 4, 469–521).

In my monograph I pushed back against this sweeping "replace the framework since it is question-begging" response throughout, but especially, and in great detail, in Chapters 5, 7, and 10. In Chapter 5 I quoted Goldman (1967) in his paper famously advancing the very first modern *externalism* (the causal theory of knowledge) as someone whom I may nevertheless read as tacitly supportive of my position rather than, say, his more-radical erstwhile philosophical ally, Hilary Kornblith. The last sentence of that famous paper is where Goldman precisely notes that his new externalist epistemology offers us a less-than-totalising, less-than-eliminativist world-view – an epistemology that is *irreflexive*:

> I think my analysis shows that the question of whether someone knows a certain proposition is, in part, a causal question, although, of course, the question of what the correct analysis is of 'S knows that *p*' is not a causal question (Goldman 1967, 372).

Of course? An irreflexive theory cannot be *reflexively incoherent* of course, but if an implicit awareness of the threat of this kind of *peritrope* was not behind the otherwise stipulative limit early Goldman placed on his theory one is left wondering what was. What Goldman, I suggest, realised, was that to generalise from a piecemeal alternative theory of knowledge to an entire, overarching normative epistemic framework (including *justification*, *rationality*, the *reflexive status of the philosopher advancing and evaluating said theory*…) needs a lot more philosophical and metaphilosophical work than is gestured towards or acknowledged by "only construct it, and move in".

REFERENCES

Alston, W. 1985. Concepts of epistemic justification. *Monist*, 68: 57-89, reprinted in Alston, 1989b, 81-115.

Bergmann, M. 2006. *Justification Without Awareness*. Oxford: Clarendon Press.

Churchland, P. M. 1993. Evaluating our self-conception. *Mind & Language*, 8: 211-222.

Foley, R. 1993. *Working Without a Net: A Study of Egocentric Epistemology.* New York and Oxford: Oxford University Press.

Goldman, A. 1967. A causal theory of knowing. *Journal of Philosophy,* 64: 357-372.

Lockie, R. 2018. *Free Will and Epistemology: A Defence of the Transcendental Argument for Freedom*. London and New York: Bloomsbury Academic.

Lockie, R. 2003. Transcendental arguments against eliminativism. *British Journal for the Philosophy of Science*, 54: 569-589.

Lucretius. 1947. *De Rerum Natura*, vol. III; tr. and commentary, Cyril Bailey, Oxford: Oxford University Press.

Plantinga, A. 1993. *Warrant: The Current Debate,* Oxford: Oxford University Press.

EuJAP | Vol. 15 | No. 2 | 2019
UDK: 123:165.322
https://doi.org/10.31820/ejap.15.2.1

TO BE ABLE TO, OR TO BE ABLE NOT TO? THAT IS THE QUESTION. A PROBLEM FOR THE TRANSCENDENTAL ARGUMENT FOR FREE WILL

NADINE ELZEIN
University of Oxford, Lady Margaret Hall

TUOMAS K. PERNU
University of Helsinki and King's College London

Original scientific article – Received: 05/06/2019 Accepted: 09/10/2019

ABSTRACT

A type of transcendental argument for libertarian free will maintains that if acting freely requires the availability of alternative possibilities, and determinism holds, then one is not justified in asserting that there is no free will. More precisely: if an agent A is to be justified in asserting a proposition P (e.g. "there is no free will"), then A must also be able to assert not-P. Thus, if A is unable to assert not-P, due to determinism, then A is not justified in asserting P. While such arguments often appeal to principles with wide appeal, such as the principle that 'ought' implies 'can', they also require a commitment to principles that seem far less compelling, e.g. the principle that 'ought' implies 'able not to' or the principle that having an obligation entails being responsible. It is argued here that these further principles are dubious, and that it will be difficult to construct a valid transcendental argument without them.

Keywords*: Determinism, epistemic deontologism, free will, libertarianism, normativity, 'ought' implies 'able not to', 'ought' implies 'can', PAP, practical deontologism, reasons, responsibility, transcendental arguments*

1. Introduction

Transcendental arguments are typically aimed at refuting sceptical positions. What is distinctive about transcendental arguments is that they do not seek to challenge the sceptic's premises directly. Rather, they might proceed in one of two ways:

Firstly, a relatively modest form of transcendental argument may begin with some fact *x* that is taken to be uncontroversial or obvious (enough so that even the sceptic cannot escape being committed to it) and by arguing that the sceptic's position is inconsistent with *x.* On this view, the sceptic's argument is not self-refuting, but the sceptic's own commitments cannot be rendered consistent with her conclusion.

Secondly, a more ambitious form of transcendental argument seeks to establish that the sceptic's stance is self-refuting, as opposed to merely being inconsistent with independently inescapable commitments. In this case, the argument will proceed first by identifying some fact *x* that is argued to be a necessary condition of the very possibility of the sceptic being able to assert her argument, and then by showing that the sceptic's conclusion cannot possibly be true consistent with *x*. Thus, if the sceptic is able to put forward an argument at all, the argument will be self-refuting. The sceptic essentially proves her own conclusion false the moment she asserts it.

Our aim is to pinpoint and assess some of the key commitments involved in constructing arguments of this sort, with a particular focus on ambitious transcendental arguments in favour of a libertarian stance in the free will debate. We maintain that the success of these arguments depends on whether we can defend not only the compelling principles that typically make these arguments appealing, but also some more dubious principles; those connecting our capacity to make rational choices not only with our ability to do so, but also with our ability to *avoid* doing so.

2. Transcendental Argument

Transcendental arguments are traditionally most strongly associated with Kant, who used the method to argue (primarily targeting Hume) that *a priori* concepts can be legitimately applied to objects of our experience, and to argue (primarily targeting Cartesian scepticism) against idealism (Kant, 1998/1781). Since Kant, the general method has commonly been

associated with responses to external world scepticism in epistemology.[1] It's rarer for this argumentative strategy to be invoked in relation to free will, although Kant's own work on free will certainly has echoes of this strategy, and there have been at least a handful of other notable examples. As far back as ancient Greece, Epicurus argues as follows:

> He who says that all things happen of necessity can hardly find fault with one who denies that all happens by necessity; for on his own theory the argument is voiced by necessity (Epicurus, 1964: fragment XL).

Epicurus does not make it entirely clear why an argument that is voiced by necessity could not be a valid argument for all that. Presumably, the driving assumption is that an argument voiced by necessity is not voiced freely, but he does not clearly spell out why this is taken to undermine the conclusion of the argument. There are, however, a number of ways in which this stance might be motivated.

While not usually regarded as an example of a transcendental argument, Kant's own reasoning in relation to free will in final section of the *Groundwork* (1997/1785) and in the *Critique of Practical Reason* (1997/1788) suggests, among other things, that one must presuppose one's own freedom in order to practically act in the pursuit of rational ends. For instance, he argues:

> Now, one cannot possibly think of a reason that would consciously receive direction from any other quarter with respect to its judgements, since the subject would then attribute the determination of his judgement not to reason but to an impulse. Reason must regard itself as the author of its principles independently of alien influences; consequently, as practical reason or the will of a rational being it must regard itself as free, that is, the will of such being cannot be a will of his own except under the idea of freedom, and such a will must in a practical respect thus be attributed to every rational being (Kant 1997/1785).

If we must presuppose our own freedom in order to act rationally, then, according to Kant, a commitment to free will is inescapable for any rational being. Moreover, if Kant is right to suppose that we cannot act rationally without presupposing that we have freedom of the sort that would be

[1] Most influentially, by Strawson (1966), but see also Putnam (1981), Peacocke (1989), Cassam (1999), and Stern (1998).

incompatible with determinism, then it seems to follow that it's also an essential precondition of choosing to argue in favour of a sceptical outlook, at least insofar as one takes oneself to have any practical reason for doing so.[2]

While Kant's argument explicitly draws on worries about *practical* normativity, the Epicurean point could just as easily rest on worries about *epistemic* normativity. In the latter case, it will be our justification for believing or asserting a conclusion, rather than our justification for acting more broadly, which is taken to commit us to supposing ourselves to be free. Insofar as our status as either practically or theoretically rational entails a certain sort of responsiveness to normative pressures, and insofar as this can be linked with a libertarian understanding of freedom, either might provide a fruitful basis for a suitable transcendental argument for such freedom.

More recently, Lockie (2018) has provided a number of detailed transcendental arguments for libertarianism, which draw on theorising about the relation between freedom, duty, and epistemic normativity, in order to show that any attempt to argue in favour of a deterministic or sceptical position must be self-refuting.

Lockie's argument rests on the idea that freedom is an essential component of epistemic justification. He also draws on the Kantian principle that 'ought' implies 'can' in order to show that determinism poses a serious threat to our capacity to respond intelligibly to epistemic norms. Hence indeterminism is taken to be a necessary prerequisite of anyone being able to justifiably reason to a conclusion – including the conclusion that determinism is true. This requires a broadly internalist and deontological conception of epistemology, according to which the ability to responsibly meet our epistemic duties is a necessary component of epistemic justification (see especially, Lockie 2018, 7-26). If determinism robs us of this ability, then it also robs us of the ability to justify a deterministic conclusion. Hence Lockie's argument forms the basis for an ambitious transcendental argument in favour of libertarian free will.

There is also scope for more modest transcendental arguments, which rest on worries about the practical feasibility of free will scepticism. It has recently been suggested that we ought to interpret Strawson's famous

[2] This has some clear parallels with Korsgaard's explicitly transcendental argument in favour of recognising moral obligations towards others, where valuing our own practical identity is taken to be necessary for having any practical reasons at all, and this is taken to commit us to recognising the value of others' rational nature on parallel grounds to the way that we must, inescapably, value our own rational nature too (Korsgaard 1996).

argument in *Freedom and Resentment* (1962) as a form of transcendental argument for compatibilism (Pereboom 2016; Coates 2017). Essentially, Strawson doubts that we can take free will scepticism seriously, given the commitments that come with the practical perspective forced upon us by our nature as practical agents. It is hardly unintelligible, on this account, to assert that we lack free will, but it may nonetheless be a practical impossibility to wholehearted maintain this view full time.

For the purposes of this discussion, we will put the Strawsonian argument for compatibilism to one side and focus solely on ambitious versions of the transcendental argument for libertarianism; on the question of whether we might have reason to suppose that arguments in favour of determinism are self-refuting in some way. The point is explicit (though underdeveloped) in Epicurus's argument, and is merely hinted at in Kant's reasoning, though it is developed thoroughly and explicitly by Lockie.

Insofar as there is a common theme here, however, the essential claims from which the argument is variously constructed appear to be something like the following:

1. 'Ought' implies 'can' (OIC).
2. Actualism about alternative possibilities: That is, the thesis that determinism rules out the ability to do otherwise; alternative possibilities of the sort required for the ability to do otherwise must be available as things actually are, holding the past and the laws of nature constant (AAP).
3. The ability to do otherwise is a necessary condition of responsibility (PAP).

The Kantian and the Lockiean arguments invoke different further principles pertaining to the sort of normative pressure required for rational action or assertion, while the Epicurean argument leaves this unstated. Though presumably, for Epicurus too, there must be some implicit assumption about the rational requirements for asserting a thesis, where it is supposed that determinism might plausibly preclude us from meeting those requirements. The Kantian principle seems to be something like the following:

4. In order to have any reason to do anything at all, we must have the ability to respond rationally to practical norms (PD).

Let's call this thesis Practical Deontologism. In contrast, the principle that Lockie's argument invokes is explicitly related to epistemic duty:

5. In order to be justified in making any assertion, we must have the ability to respond rationally to epistemic norms (ED).

Lockie calls this thesis Epistemic Deontologism. Either 4 (PD) or 5 (ED) may feasibly be invoked, alongside all or some subset of claims along the lines of 1-3, in an ambitious transcendental argument for libertarian free will. These are all claims that we will be happy to grant, at least for the sake of this discussion. Although they are all controversial, they also each seem to have a fair degree of independent plausibility.

However, we hope to show that in order for any argument of this sort to succeed, there must also be a commitment to one of the following further claims, which we take to be significantly more controversial than the others:

6. 'Ought' implies 'able not to' (OIANT).
7. Duty entails responsibility; no one ought to do something unless they would be responsible for doing it (DER).

Note, that if we take the truth of PAP for granted, these claims essentially become equivalent: The basic idea is that in order to be obligated to do *x*, we either directly need the ability to refrain from doing *x*, or we need to be responsible for doing *x*, where that, in turn, entails (given PAP) an ability to refrain from doing *x.* Hence what will be needed, in relation to meeting our practical or epistemic obligations, is not merely to be able to, but also to be able not to. That is, for this argumentative strategy to be effective, there are negative and positive preconditions of justifiably acting, asserting, or believing; not only must we be capable of doing what we *ought* to do, but we must also be capable of *not* doing what we ought to. It is this aspect of the argument that we take to be problematic.

3. Determinism, Alternatives, and 'Ought' Implies 'Can'

3.1. Determinism and AAP

Following Van Inwagen, we may define determinism as the conjunction of the following two theses:

a) For every instant of time, there is a proposition that expresses the state of the world at that instant.

b) If *p* and *q* are any propositions that express the state of the world at some instants, then the conjunction of *p* with the laws of nature entails *q*.[3]

If determinism is true, only one future course of events will be possible, consistent with holding fixed the laws of nature and the way that things were in the past. While it might seem intuitive to suppose, at first sight, that the truth of this thesis rules out the ability to do otherwise, there is a great deal of controversy surrounding this point.

According to one reading – we call this the 'actualist' reading[4] – an agent is only able to do otherwise, in the relevant sense, if she is able to do otherwise as things actually stand, holding the past and the laws of nature constant (AAP). On this actualist understanding, determinism rules out alternative possibilities.[5] In contrast, many theorists favour a counterfactual or dispositional reading. On the counterfactual reading, an agent is able to do otherwise if, for instance, she would have done otherwise had she chosen to.[6] On a dispositional reading, an agent could have done otherwise if she would have done otherwise had she been placed in different circumstances.[7] Determinism is consistent with the ability to do otherwise in both of these senses.

While AAP is controversial within the free will debate, it does seem to capture at least one sense of 'able to do otherwise', which goes beyond the conditional and dispositional senses, and which many take to be important for free will. An agent who can act otherwise in the conditional and dispositional senses is one that acts deliberately, acts on the basis of her own choices, and is adequately sensitive to important features of her environment. Many philosophers suppose that this suffices to establish that she acts freely and responsibly. However, while these abilities are almost universally acknowledged to be necessary for moral responsibility, many incompatibilist philosophers have doubts about whether they are sufficient. If the agent is unable to choose otherwise, given the way things *actually are*, we may worry that she cannot really, in some crucial sense, escape acting the way that she does. E.g. we may worry that she still lacks the ability to act otherwise in a sufficiently robust sense; it may still seem

[3] See van Inwagen (1983, 65). A similar definition is given in van Inwagen (1975, 186).
[4] See Elzein and Pernu (2017).
[5] Notable defences of the actualist analysis include Campbell (1951), Chisholm (1964), Lehrer (1968), van Inwagen (1983; 2000; 2004; 2008), and Kane (1999).
[6] Notable defences of the counterfactual analysis include Moore (1903), Ayer (1954), Smart (1961), Schlick (1939), Lewis (1981), and Berofsky (2002).
[7] Notable defences of the dispositional analysis include Fara (2008), Smith (1997; 2003) Vihvelin (2004; 2011; 2014).

unfair to blame her for what she does if she could not *actually* escape blame, given the way things are. In any case, we will grant AAP for the purposes of this discussion.

3.2. Obligation and OIC

The principle that 'ought' implies 'can' (OIC) is popular,[8] but nonetheless remains controversial.[9] There is, however, undoubtedly a great deal of intuitive appeal in the idea that there is something wrong with supposing that demands can be placed on an agent which are impossible for that agent to meet.

In order for this principle to be utilised effectively in any transcendental argument for free will, however, we will need to say something about the sense of 'can' invoked by the principle. Specifically, we will need to suppose that the principle is convincing even granted an actualist reading of 'can'. That is, we must suppose that an agent cannot be obligated to do something unless that agent is able to do it, as things actually stand, holding the past and the laws of nature constant. If we wish to show that determinism undermines our ability to do what we ought to do, in the sense relevant to OIC, then we had better suppose that this pertains to the same sense of 'able to' according to which determinism might plausibly be thought to rob us of the ability to do otherwise.

For the sake of this discussion, we will grant both OIC, and that the sense of 'able to do otherwise' that is relevant to OIC is that invoked by AAP. That is, we will grant that determinism rules out alternative possibilities, and that it does so in a way that entails that we are unable to do otherwise, which, in conjunction with OIC, entails that we cannot be *obligated* to do otherwise.

---

[8] The principle is commonly thought to originate with Kant (1998/1781; 2017/1797; 1998/1793; 1996/1793), and was famously defended by Moore (1922). Since then it is more often taken to be a basic platitude than explicitly argued for, but there are some explicit defences of the principle. See Sapontzis (1991), Griffin (1992), Streumer (2003; 2007; 2010), and Vranas (2007). For defences of related principles, see Graham (2011) and Kühler (2013).

[9] Notable critiques include Lemmon (1962), Williams (1965), Brouwer (1969), Trigg (1971), van Fraassen (1973), Brown (1977), Sinnott-Armstrong (1984; 1988), Rescher (1987, chap. 2, pp. 26-54), Saka (2000), Fischer (2003), and Heintz (2013). Cf. Kekes (1984) and Stern (2004). For empirical objections to the principle, see Semler and Henne (2019).

3.3. Normative Pressures and PAP

There are various ways in which a transcendental argument might run. It may only be necessary to appeal to our capacity to respond to normative pressures, in which case it is not obvious we need to invoke the idea of responsibility at all. But the argument could proceed via a consideration of responsibility if what is taken to be important is not merely the ability to respond to normative pressures, but the ability to be responsible for doing so. In the latter case, the argument may need to make use of PAP: The principle that alternative possibilities are a required for responsibility.

PAP has been under frequent attack at least since Frankfurt's famous attempt to refute the principle (Frankfurt 1969). For present purposes, we will accept PAP, although later we will have reason to consider whether the principle is of central importance to plausible versions of the transcendental argument.

In any case, what any version of the argument will need is some appeal to a normative principle, which bears on when we could have an intelligible basis for making an assertion or for justifying our commitment to a conclusion. Rational justifications for either belief or action must be taken to depend on some sort of ability to respond to normative pressures – whether practical or epistemic. It is this ability that will, if the argument is convincing, be threatened by determinism.

3.4. The Basic form of Transcendental Argument

Suppose that we take the principles above to be defensible. This gives us a framework for constructing an ambitious version of the transcendental argument for libertarianism. A simple argument will not rest on PAP, but will instead appeal directly to worries about our ability to respond to normative pressures. This will go as follows:

(1) If determinism is true, then nobody is able to do otherwise (from AAP).
(2) If nobody is able to do otherwise, then nobody is able to assert or conclude otherwise (uncontroversial entailment).
(3) If nobody is able to assert or conclude otherwise, then nobody ought to assert or conclude otherwise (from OIC).
(4) If nobody ought to assert or conclude otherwise, then nobody can have an adequate rational basis to assert or to justifiably conclude otherwise (from either PD or ED).

(5) If determinism is true, then nobody could have an adequate rational basis to assert or justifiably conclude otherwise (from 1-4).
(6) If determinism is true, then nobody could have an adequate rational basis for any actual assertion or conclusion (from…?)

The problem here is that (6) does not follow from the preceding steps. It certainly doesn't follow from (5) alone. In fact, we only seem entitled to (5). Clearly, however, (5) is a weaker claim than the one needed to establish that any argument for determinism is self-refuting. This would establish that the proponent of determinism cannot have any justification for asserting any *alternative* conclusion. This looks unproblematic. Insofar as one takes oneself to have a decisive rational basis for asserting a particular conclusion, it follows rather trivially that one cannot be justified in asserting the opposite conclusion instead. To render the determinist's stance problematic, we need a stronger conclusion: That the proponent of the argument for determinism cannot have any rational justification for asserting her *actual* conclusion.

This could be done either by invoking a principle linking responsibility to duty (DER) alongside PAP, or by invoking the principle that 'ought' implies 'able not to' (OIANT). In either case, we need some means of supposing that the capacity not to fulfil a duty is a necessary condition of being duty-bound, so we end up either directly or indirectly, arriving at something like OIANT.

It's fairly easy to see how the inclusion of OIANT on its own would help to establish a strong enough conclusion:

(1) If determinism is true, then nobody is able to do otherwise (from AAP).
(2) If nobody is able to do otherwise, then it follows that nobody who makes an assertion or reaches a conclusion could assert or conclude anything other than what they actually do (uncontroversial entailment).
(3) If nobody is able to assert or conclude otherwise than they actually do, then nobody ought to assert or conclude as they actually do (from OIANT).
(4) If nobody ought to assert or conclude as they actually do, then nobody can have an adequate rational basis to assert or to justifiably conclude as they actually do (from either PD or ED).
(5) If determinism is true, then nobody could have an adequate rational basis for any actual assertion or conclusion (from 1-4).

This argument *would* entail that the determinist would have no rational basis, were determinism true, on which to justify asserting or concluding anything – including the claim that determinism is true.

While the argument could be constructed by appeal OIANT, another route to the same conclusion would arrive at something that entails OIANT, but would commit to it indirectly via PAP and DER, as follows:

(1) If determinism is true, then nobody is able to do otherwise (from AAP).
(2) If nobody is able to do otherwise, then nobody can be responsible for anything that they actually do (from PAP).
(3) If nobody is responsible for anything they actually do, then nobody can be responsible with respect to the assertions they actually make or the conclusions they actually reach (uncontroversial entailment).
(4) If nobody is responsible with respect to the assertions they actually make or the conclusions they actually reach, then nobody ought to make the assertions they make or reach the conclusions that they reach (from DER).
(5) If nobody ought to make the assertions that they make or reach the conclusions that they reach, then nobody can have any rational justification for their conclusions or assertions (from PD or ED).
(6) If determinism is true, then nobody can have any rational justification for their conclusions or assertions (from 1-5).

The argument may then invoke either OIANT or else PAP alongside DER. The problem, however, is that neither OIANT nor DER are plausible. When there is a compelling practical reason for doing something or a compelling epistemic reason for believing something, we will argue that these pressures are typically independent both of whether we can *avoid* responding to the pressure and of whether we would be *responsible* for responding to the pressure. That is, practical and epistemic normative pressures involve the ability to respond to our actual reasons or our actual evidence, and rely neither on our ability to avoid responding to these, nor on whether we would be responsible for responding.

We might suppose that an epistemically rational agent aims to have beliefs that "track" the truth and that a practically rational agent aims to make choices that that "track" their reasons for action.[10] If normative pressures

[10] While the former idea is notably associated with (Nozick 1981) and the latter view is associated with Fischer and Ravizza (1998; see, also, Fischer 1987), the claim being made

are understood in terms of the obligation to make our assertions and conclusions, as far as possible, track what there is *reason* to assert or to conclude, it's not at all obvious that either responsibility or the ability to assert or conclude otherwise should be relevant to these pressures at all. While an agent's lack of freedom or responsibility with respect to these pressures may well have an important bearing on whether they can intelligibly be held accountable for their beliefs or assertions, they will *not* obviously have any parallel bearing on the strength of the agent's reasons for asserting or believing what they do.

4. The Implausibility of OIANT and DER

4.1. The Problem with OIANT

While OIC might seem highly intuitive, OIANT appears to be far less so. While some argue that the two principle are symmetrical in such a way that we ought to accept one so long as we accept the other (e.g. Haji 2002, see especially page 29), it has also been noted that the alleged symmetry is hardly obvious, and unlike OIC, OIANT is rarely seen as similarly axiomatic (Nelkin 2011, 102). Moreover, we might suppose that there is an intuitive rationale for endorsing OIC that simply does not apply to OIANT; we maintain that OIC is plausible because it seems *unreasonably demanding* to insist that anyone ought to do the impossible. The fact that something is unavoidable, in contrast, certainly does not entail that it would be unreasonably demanding to suppose that someone ought to do it.

Moreover, whether we focus on the epistemic or the practical realm (e.g. on the moral or on the prudential), we will easily find cases in which this principle appears highly counterintuitive. For instance, suppose that you are unable to put your hand into a flame and hold it there for five minutes. Does this really plausibly entail that it's false that you ought to *avoid* putting your hand in a flame and holding it there for five minutes? Or suppose that you are unable to avoid believing that 1 + 1 = 2. Does this entail that that you lack a strong rational justification for believing that 1 + 1 = 2? Likewise, suppose that you are unable to murder someone in cold blood. Does this plausibly entail that it's false that you ought not to murder

here is committed neither to Nozick's externalism about epistemology nor to Fischer and Ravizza's semi-compatibilism about responsibility. In relation to knowledge, the point is that a rational agent *aims* to have truth-responsive beliefs, where this may be understood as a response to an epistemic duty, consistent with the sort of internalist epistemic deontologism defended by Lockie (2018), say. And while we are suggesting that *practical rationality* requires the ability to respond to reasons, we are not arguing, as Fischer and Ravizza do, that this suffices for moral responsibility.

anyone cold blood? In all of these cases, it seems plausible to suppose that the answer is no.

The reasons for this cut to what is distinctive about normative pressures. Perhaps you cannot put your hand into a flame and hold it there for five minutes, but this is hardly relevant to the reasons why you ought not to do such a thing. You ought not to do it because you have a very strong prudential *interest* in avoiding unnecessary pain and injury. This prudential interest will still exist regardless of whether you cannot *help* but avoid it. Similarly, your reason for believing that 1 + 1 = 2 seems to be just as strong regardless of whether you have the ability to doubt it. The reason is provided by the strength of the mathematical case in favour of concluding that 1 + 1 = 2; that is the strength of the evidence you have on the basis of which to suppose it's *true.* Likewise, your reasons for not murdering someone in cold blood are based on the fact that it would be morally *wrong*, not on the fact that you are able to do it.

If we accept OIC, this entails that it would be false that someone who is incapable of avoiding putting their hand in a flame ought not to do so. The obvious rationale is that it cannot be a good idea to do something if that something is literally impossible to do. The practical plausibility of this view appears to be grounded in the fact that it's never practically a good idea to attempt the impossible.

Our point, however, is that it may well be worth attempting the inevitable, especially if there is a causal link between your attempt and your success in that attempt. It may well be inevitable that the moment you realise your hand is in the flame, you retract it fairly quickly. But this doesn't obviously entail that doing so is not also a *good idea*. You have strong reasons to do it based on the fact that it's in your interests and you are easily capable of doing it. Similarly, if we accept OIC, a person who is *incapable* of believing that 1 + 1 = 2 is not a person who *ought* to believe that 1 + 1 = 2. But this does not entail that a person who cannot help but believe it has no reason to believe it.

A plausible form of epistemic deontologism will entail that we have a duty to believe what there is strong evidence for believing, insofar as we are capable of understanding and accurately assessing that evidence. There is no obvious parallel for supposing that we also need the ability to *doubt* what there is overwhelming evidence to believe. In the case of simple mathematical truths, most of us are likely to find these fairly indubitable. But it seems odd, to say the least, that we should suppose (in a stark reversal of the Cartesian approach!) that a truth's status as indubitable actually positively *undermines* our justification for believing it.

One worry may be that we must be committed, in principle, to a strong parallel between 'ought' and 'ought not' requirements. For instance, Lockie argues that because

> determinism globally denies us the negative, 'irrational', 'unjustified' aspects of any internalist value terms, it removes from us the ability to distinguish and use the positive 'rational', 'justified' aspects of such terms. (If one affects to make no sense of anything being *not red*, one cannot distinguish and use the predicate *red*). (Lockie 2018, 182)

It is precisely this principle, however, that we take issue with.

Firstly, the parallel between supposing there are no unjustified beliefs or actions, on the one hand, and "affecting to make no sense of redness", on the other, is dubious: The claim is not that we can *make no sense* of any belief being unjustified, but that if determinism should turn out to be true, then as a matter of fact, nobody is under an obligation not to hold the beliefs they have or under an obligation not to make the assertions that they do. There is an important difference. Consider the idea of non-existence; it is a simple tautology that there exist no things that don't exist. But we can understand the *concept* of non-existence even if, as a matter of fact, there are no things that don't exist. We are able to make sense of the concept because we are able to think in modal terms; we can contemplate hypothetical scenarios.

There is a great deal of disagreement regarding whether or not determinism is true. Even if we suppose that determinism is true, and we embrace something like OIC, it's not at all obvious that we should be unable to make any sense of the idea that some people ought to believe or assert something different to what they do. This requires that we can *imagine* a world in which determinism is false, and can think about the obligations we would be under in such a world. This is perfectly consistent with supposing that, as a matter of fact, nobody has such obligations as things actually are.[11]

More importantly, the relevant discrimination capacities do not seem to have been located in quite the right place: the normative pressure comes

[11] Compare the point here with a somewhat parallel argument, which Stroud (1968) makes in response to epistemic versions of the transcendental argument: Is it obvious that we need there to be objects in the external world in order to make sense of our experiences? Perhaps all we need is to have the *impression* or the *belief* that there are. Much the same seems to be true with respect to irrational or unjustified beliefs.

from the strength of the evidence. The relevant ability involves being able to discriminate between *strong evidence* and *weak evidence.* A person may well have the ability to discriminate between strong and weak evidence, even if they are not capable of believing anything on the basis of weak evidence, or of doubting something for which there is overwhelmingly strong evidence.

4.2. The Problem with DER

We maintain that there is a parallel issue with DER, the principle that in order to have a reason to do or believe something, we would have to be held responsible for doing or believing it.

Again, this appears to misplace the source of the relevant normative pressures. The reasons we have to believe something are dependent on the strength of the evidence in its favour; not on the epistemic agent's blameworthiness or praiseworthiness for so believing. Similarly, the practical reasons we have for acting depend on the moral or prudential case in favour of so acting. Again, where there are strong reasons to do something, these reasons are typically not dependent on whether an agent would be praiseworthy for doing it or blameworthy for failing to do it.

We are not arguing that claims about whether an agent is morally or epistemically blameworthy or praiseworthy are entirely independent of the agent's moral or epistemic reasons: It is clear that if anyone is *ever* epistemically praiseworthy, a necessary precondition of this is that the agent has good evidence on the basis of which she arrives at her belief. Similarly, if anyone is ever morally or prudentially praiseworthy, a necessary precondition of this is that she had good reasons on the basis of which to act as she did. What we deny, however, is that there is any entailment in the opposite direction: that is, that being praiseworthy is a precondition of having good moral or epistemic reasons. Praiseworthiness, if there is such a thing, *depends* on there being independent sources of epistemic and practical normativity, not *vice-versa*.

For one thing, is seems that agents may not be sophisticated enough to be held responsible for their beliefs and actions but may nonetheless have *reasons* for those beliefs and actions. Consider a five-year-old child who refrains from playing with the loose electrical cables coming out of a live plug socket on the basis that a parent has told her not to. Plausibly, the child is not responsible for her actions since she doesn't really appreciate the reasons why she ought not to play with the electrical cables. But plausibly she ought *not* to play with them. When her parents tell her that she ought not to touch that live wire, they can hardly be accused of lying

to her. She ought not to touch that live wire. The reason why she ought not to touch the wire is certainly *not* that she will be praiseworthy if she avoids touching it and blameworthy if she touches it (neither of *those* claims seems plausibly true). In fact, her responsibility doesn't come into it. Rather, she ought not to touch it because she's likely to receive a nasty electric shock if she does.

With respect to epistemic reasons, it seems even more clear that the normative pressures arising from the strength of evidence are not in any way derived from the agent's status as responsible. Suppose the five-year-old works out that 5 x 5 = 25. Perhaps this is quite a difficult calculation for a child or her age and abilities, and it would therefore be unreasonable to suppose that she could be held responsible for successfully working it out. It would certainly be unreasonable to blame her for getting it wrong. None of this seems to have much bearing, however, on why we might suppose that she *ought* to believe that 5 x 5 = 25. She ought to believe it because it's *true* and because it's strongly supported by mathematical logic.

Again, the point is that normative pressures arise from facts about what there is evidence to believe and what there is reason to do. These facts do not depend on whether we are responsible. The norms that govern rational belief and behaviour are independent of considerations about whether anyone is responsible for their beliefs and actions.

5. Conclusion

While there are a number of plausible principles underpinning transcendental arguments for the freedom of the will, they also appear to rest, inevitably, on some principles that we may have good grounds for rejecting. Even if our duties rest on our ability to fulfil them, it is not at all obvious that they similarly rest on any parallel ability *not* to fulfil them. And while we may have reasons to suppose that our responsibility in relation to our beliefs and actions depends on our reasons, it is far from obvious that there is any dependence in the other direction. It seems, then, that if any transcendental argument in favour of free will is to succeed, it will have to be a significantly more modest form of argument than the sort we have been considering here. It is difficult to see why the determinist could not have good reasons to assert her position without risk of contradiction or self-refutation.

REFERENCES

Ayer, A. J. 1954. Freedom and necessity. In his *Philosophical Essays*, 271-284. New York: St Martin's Press.

Berofsky, B. 2002. Ifs, cans, and free will: the issues. In *The Oxford Handbook of Free Will*, ed. Robert Kane, 181-201. Oxford: Oxford University Press.

Brouwer, F. E. 1969. A difficulty with 'ought implies can'. *The Southern Journal of Philosophy* 7: 45-50.

Brown, J. 1977. Moral theory and the ought–can principle. *Mind* 86: 206-223.

Campbell, C. A. 1951. Is 'freewill' a pseudo-problem? *Mind* 60: 441-465.

Cassam, Q. 1999. *Self and World.* Oxford: Oxford University Press.

Chisholm, R. 1964. Human freedom and the self. *The Lindley Lecture.* Lawrence, KS: Department of Philosophy, University of Kansas, 3-15. Reprinted in *Free Will*, second edition, edited by Gary Watson, 26-37. Oxford: Oxford University Press. 2003.

Coates, D. J. 2017. Strawson's modest transcendental argument. *British Journal for the History of Philosophy* 25: 799-822.

Elzein, N., and T. K. Pernu 2017. Supervenient freedom and the free will deadlock. *Disputatio* 9: 219-43.

Epicurus. 1964. *Letters, Principal Doctrines and Vatican Sayings.* Translated by Russel. M. Geer. Indianapolis and New York: Bobbs-Merril.

Fara, M. 2008. Masked abilities and compatibilism. *Mind* 117: 843-865.

Fischer, J. M. 1987. Responsiveness and moral responsibility. In *Responsibility, Character, and the Emotions*, ed. F. Schoeman, 81-106. Cambridge: Cambridge University Press.

Fischer, J. M., and M. Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Fischer, J. M. 2003. 'Ought-implies-can', causal determinism and moral responsibility. *Analysis* 63: 244-250.

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

Graham, P. A. 2011. 'Ought' and ability. *Philosophical Review* 120: 337-382.

Griffin, J. 1992. The human good and the ambitions of consequentialism. *Social Philosophy and Policy* 9: 118-132.

Haji, I. 2002. *Deontic Morality and Control*. Cambridge: Cambridge University Press.

Heintz, L. L. 2013. Excuses and "ought" implies "can". *Canadian Journal of Philosophy* 5: 449-462.

Kane, R. 1999. Responsibility, luck, and chance: reflections on free will and indeterminism. *The Journal of Philosophy* 96: 217-240.

Kant, I. 1998/1781. *The Critique of Pure Reason.* Translated and edited by P. Guyer and A. W. Wood. Cambridge: Cambridge University Press.

Kant, I. 1997/1785. *Groundwork of the Metaphysics of Morals.* Edited and translated by Mary Gregor. Cambridge: Cambridge University Press.

Kant, I. 1977/1788. *Critique of Practical Reason.* Translated and edited by Mary Gregor. Cambridge: Cambridge University Press.

Kant, I. 2017/1797. *The Metaphysics of Morals.* Edited by Lara Denis and translated by Paul Guyer and Mary Gregor. Cambridge University Press.

Kant, I. 1998/1793. Religion within the boundaries of mere reason. In *Religion Within the Boundaries of Mere Reason and Other Writings*, translated and edited by Allen Wood and George Di Giovanni, 31-191. Cambridge: Cambridge University Press.

Kant, I. 1996/1793. On the common saying: That may be correct in theory, but it is of no use in practice. *Practical Philosophy*, edited by Mary Gregor. Cambridge: Cambridge University Press.

Kekes, J. 1984. 'Ought implies can' and two kinds of morality. *Philosophical Quarterly* 34: 459-467.

Korsgaard, C. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.

Kühler, M. 2013. Who am I to uphold unrealizable normative claims? *Autonomy and the Self*, eds. M. Kühler and N. Jelinek, 191-209. Dordrecht: Springer.

Lehrer, K. 1968. Cans without ifs. *Analysis* 29: 29-32.

Lemmon, E. J. 1962. Moral dilemmas. *The Philosophical Review* 70: 139-158.

Lewis, D. 1981. Are we free to break the laws? *Theoria* 47: 113-121.

Lockie, R. 2018. *Free Will and Epistemology.* London: Bloomsbury.

Moore, G. E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

Moore, G. E. 1922. The nature of moral philosophy. In his *Philosophical Studies*. Routledge and Kegan Paul.

Nelkin, D. K. 2011. *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.

Nozick, R. 1981. Knowledge and scepticism. In his *Philosophical Explanations*. Cambridge, MA: Harvard University Press.

Peacocke, C. 1989. *Transcendental Arguments in the Theory of Content.* Oxford: Oxford University Press.

Pereboom, D. 2016. Transcendental arguments. In *The Oxford Handbook of Philosophical Methodology,* eds. H. Cappelen, T. Szabó

Gendler, and J. Hawthorne, 444-62. Oxford: Oxford University Press.
Putnam, H. 1981. *Reason, Truth, and History.* Cambridge: Cambridge University Press.
Rescher, N. 1987. *Ethical Idealism: An Inquiry into the Nature and Function of Ideals*. Berkeley: University of California Press.
Saka, P. 2000. Ought does not imply can. *American Philosophical Quarterly* 37: 93-105.
Sapontzis S. F. 1991. 'Ought' does imply 'can'. *The Southern Journal of Philosophy* 29: 382-393.
Schlick, M. 1939. When is a man responsible? In his *Problems of Ethics,* 143-156. New York: Prentice-Hall.
Semler, J., and P. Henne. 2019. Recent experimental work on 'ought' implies 'can'. *Philosophy Compass* e12619.
Sinnott-Armstrong, W. 1984. 'Ought' conversationally implies 'can'. *Philosophical Review* 93: 249-261.
Smart, J. J. C. 1963. Free will, praise and blame. *Mind* 70: 291-306.
Smith, M. 1997. A theory of freedom and responsibility. In *Ethics and Practical Reason*, eds. G. Cullity and B. Gaut, 293-317. Oxford: Oxford University Press. Reprinted in his *Ethics and the A Priori.* Cambridge: Cambridge University Press, 2004.
Smith, M. 2003. Rational capacities, or: how to distinguish recklessness, weakness, and compulsion. In *Weakness of Will and Practical Irrationality*, eds. S. Stroud and C. Tappolet, 17-38. Oxford: Clarendon Press. Reprinted in his *Ethics and the A Priori.* Cambridge: Cambridge University Press, 2004.
Strawson, P. F. 1962. Freedom and resentment *Proceedings of the British Academy* 48: 1-25.
Strawson, P. F. 1966. *The Bounds of Sense: An Essay on Kant's Critique of Pure Reason*. London: Methuen.
Stern, R. ed. 1998. *Transcendental Arguments*. Oxford: Oxford University Press.
Stern, R. 2004. Does 'ought' imply 'can'? And did Kant think it does? *Utilitas* 16: 42-61.
Streumer, B. 2003. Does 'ought' conversationally implicate 'can'? *European Journal of Philosophy* 11: 219-228.
Streumer, B. 2007. Reasons and impossibility. *Philosophical Studies* 136: 351-384.
Streumer, B. 2010. Reasons, impossibility and efficient steps: Reply to Heuer. *Philosophical Studies* 151: 79-86.
Stroud, B. 1968. Transcendental arguments. *Journal of Philosophy* 65: 241-256.
Trigg, R. 1971. Moral conflict. *Mind* 80: 41-55.

van Fraassen, B. 1973. Values and the heart's command. *The Journal of Philosophy* 70: 5-19.
van Inwagen, P. 1975. The incompatibility of free will and determinism. *Philosophical Studies* 27: 185-199.
van Inwagen, P. 1983. *An Essay on Free Will.* Oxford: Oxford University Press.
van Inwagen, P. 2000. Free will remains a mystery. *Philosophical Perspectives* 14: 1–20.
van Inwagen, P. 2004. Freedom to break the laws. *Midwest Studies in Philosophy* 28: 336–350.
van Inwagen, P. 2008. How to think about the problem of free will. *The Journal of Ethics* 12: 327-341.
Vihvelin, K. 2004. Free will demystified: A dispositional account. *Philosophical Topics* 32: 427-450.
Vihvelin, K. 2011. How to think about the free will/determinism problem. In *Carving Nature at its Joints,* eds. J. K. Campbell and M. O'Rourke, 313-340. Cambridge, MA: MIT Press.
Vihvelin, K. 2013. *Causes, Laws, and Free Will: Why Determinism Doesn't Matter.* New York: Oxford University Press.
Vranas P. B. M. 2007. I ought, therefore I can. *Philosophical Studies* 136: 167-216.
Williams, B. 1965. Ethical consistency. *Proceedings of the Aristotelian Society*, supplementary volumes 39: 103-124.

EuJAP | Vol. 15 | No. 2 | 2019
UDK: 165.322.123.1
https://doi.org/10.31820/ejap.15.2.2

DETERMINISM AND JUDGMENT. A CRITIQUE OF THE INDIRECT EPISTEMIC TRANSCENDENTAL ARGUMENT FOR FREEDOM

Luca Zanetti

Original scientific article – Received: 30/05/2019 Accepted: 09/08/2019

ABSTRACT

In a recent book entitled Free Will and Epistemology. A Defence of the Transcendental Argument for Freedom, Robert Lockie argues that the belief in determinism is self-defeating. Lockie's argument hinges on the contention that we are bound to assess whether our beliefs are justified by relying on an internalist deontological conception of justification. However, the determinist denies the existence of the free will that is required in order to form justified beliefs according to such deontological conception of justification. As a result, by the determinist's own lights, the very belief in determinism cannot count as justified. On this ground Lockie argues that we are bound to act and believe on the presupposition that we are free. In this paper I discuss and reject Lockie's transcendental argument for freedom. Lockie's argument relies on the assumption that in judging that determinism is true the determinist is committed to take it that there are epistemic obligations – e.g., the obligation to believe that determinism is true, or the obligation to aim to believe the truth about determinism. I argue that this assumption rests on a wrong conception of the interplay between judgments and commitments.

***Keywords**: Epistemic deontology, free will, transcendental arguments, judgment*

1. Introduction

There is a long and rich tradition of arguments attempting to show that to believe in determinism is somehow self-defeating or self-refuting.[1] These arguments articulate in various ways the insight expressed by Epicurus in this oft-quoted passage:

> The man who says that all things come to pass by necessity cannot criticize one who denies that all things come to pass by necessity; for he admits that this too happens of necessity. (Epicurus 1926, 113)

In a recent book entitled *Free Will and Epistemology. A Defence of the Transcendental Argument for Freedom*, Robert Lockie revives this tradition by defending his own version of the Epicurean argument against determinism.[2]

Lockie's argument is called the 'indirect epistemic transcendental argument for freedom' (hereafter 'IETAF'). IETAF hinges on the contention that we are bound to assess whether our beliefs are justified by relying on an internalist deontological conception of justification.[3] However, the determinist denies the existence of the free will which is required in order to form justified beliefs according to such deontological conception of justification. As a result, by the determinist's own lights, the very belief in determinism cannot count as justified.

This argument doesn't prove that determinism is false. Rather, it shows that a determinist can't hold her view in a coherent manner. On this ground Lockie argues for a view he calls *presuppositional incompatibilism*, i.e., the view that we are bound to act and believe on the presupposition that we possess the kind of freedom defended by incompatibilists, namely the kind of freedom that is needed in order to fulfil our epistemic obligations.

---

[1] See Jordan (1969, 48) for a list of defenders of epicurean arguments before 1969. See Knaster (1986) for a list of defenders of epicurean arguments before 1986. Recent influential discussions include Honderich (1990a; 1990b) and, most recently, Slagle (2016) and Lockie (2018). See Honderich (1990a, 361ff) for eight different versions of the Epicurean argument from self-defeat. The indirect epistemic transcendental argument for freedom is only one of the many arguments that took inspiration from Epicurus's quote.

[2] Lockie's argument is similar in many respects to the one defended in Boyle, Grisez, and Tollefsen (1976).

[3] Throughout this paper I am concerned with *epistemic* justification only, and with deontological views that countenance the existence of *epistemic* obligations only.

The aim of this paper is to explain why IETAF fails to deliver the conclusion that determinism is self-defeating.[4]

In §2 I clarify Lockie's IETAF and argue that it relies on the following crucial contention:

Determinist's Commitment to Epistemic Obligations (Commitment$_{EO}$): In judging that determinism is true the determinist is thereby committed to take it that there are epistemic obligations.

In §3 I discuss Lockie's own preferred version of internalist epistemic deontologism, i.e., the view that justification is to be understood in terms of the fulfilment of one's perceived epistemic obligations. Crucially, Lockie endorses doxastic involuntarism – i.e., the claim that we have no direct voluntary control over the formation of our beliefs – and on this ground he is bound to reject the existence of doxastic obligations, that is, epistemic obligations about what to believe. However, he argues that we possess the relevant freedom that underpins what I shall call cognitive obligations, that is, epistemic obligations concerning how to manage one's own cognitive activities in inquiry.

In §§4-5 I introduce Lockie's transcendental argument for the ineliminability of deontological appraisal (or 'ineliminability argument', hereafter 'IA'). IA is meant to play a crucial dialectical role in Lockie's defence of IETAF. However, I argue that IA doesn't provide any motivation for Commitment$_{EO}$ and that as a result the defender of IETAF is left with the burden to provide grounds for Commitment$_{EO}$.

In §§6-7 I distinguish and evaluate three different versions of Commitment$_{EO}$. I argue that they are all false, and that their prima facie plausibility, if any, might be captured by structurally analogous claims that do not involve a commitment to epistemic obligations of any sort.

In §8 I conclude by locating Lockie's IETAF within the literature on modest and ambitious transcendental arguments, and through that comparison I argue that the use of modest transcendental arguments in the free will debate is problematic.

[4] This paper elaborates some of the remarks I have made in my review of Lockie's book. See Zanetti (2019).

2. The Indirect Epistemic Transcendental Argument for Freedom

Lockie's IETAF can be summarized as follows[5] (where by 'justification$_{ID}$' and cognate expressions I refer to *epistemic* justification as understood according to an *internalist deontological* notion of justification).

P1) If determinism is true, then no-one can do otherwise.
P2) The ability to reason otherwise is necessary for someone to be held unjustified$_{ID}$.
P3) If determinism is true, then no-one may be held unjustified$_{ID}$.
P4) If no-one may be held unjustified$_{ID}$, then no-one is justified$_{ID}$ either.
P5) If no-one is ever justified$_{ID}$, then belief in determinism is not justified$_{ID}$ either.
C) If determinism is true, then belief in determinism is not justified$_{ID}$.

Lockie comments on the conclusion of the argument as follows:

> I take it that this would be a wholly unsustainable position for the determinist to be in – that the determinist simply must resist the conclusion of this argument. (Lockie 2018, 183)

But why should this conclusion trouble the determinist? After all, the determinist's worldview rejects the existence of the sort of freedom that is needed in order to underpin epistemic obligations. Moreover, the determinist can grant that the very belief in determinism is not justified$_{ID}$, and yet she can insist that her belief is justified according to other non-deontological notions of justification (whether internalist or externalist). The determinist can argue that her belief is based on good evidential grounds, and she can also claim that it has been formed through a suitably reliable belief-forming process. Thus, the argument as it stands doesn't show that determinism is self-defeating.

The argument provides grounds for concluding that determinism is self-defeating if we add the following two premises:

Monist Epistemic Deontologism (MED): a deontological conception of justification is the sole correct account of epistemic justification.

[5] See Lockie (2018, 182-183) for more details on the argument's overall structure and the motivation for the main premises.

Judgment's Commitment to Epistemic Justification (Commitment$_{EJ}$): in judging that *p* one is thereby committed to take it that there is a justification for judging that *p*.[6]

By judging that determinism is true, the determinist is committed to take it that her very judgment in determinism is justified (via Commitment$_{EJ}$); but since epistemic justification has to be understood in deontological terms only (MED), by judging that determinism is true the determinist is thereby committed to take it that her judgment in determinism is justified$_{ED}$; and yet, by judging that determinism is true, she is also committed to the claim that the judgment in determinism is not justified$_{ID}$ (via P1-P5). Thus, by judging that determinism is true the determinist is committed to incompatible commitments: that there are no justified$_{ED}$ judgments, and that judgment in determinism is justified$_{ED}$.

The argument, as it stands, has few chances to be taken seriously by contemporary participants in the free will debate and in the debate on the nature of epistemic justification. First of all, the determinist has several options to reject the argument, most of which appeal to compatibilist approaches to the problem of free will.[7] But the most highly contentious – and widely rejected – premise is MED. If MED is false, then it is open for the determinist to argue that her belief in determinism is justified according to a non-deontological notion of justification, and thus the determinist can avoid the charge of being endorsing a self-defeating standpoint. MED could be rejected either by arguing that there is no notion of epistemic justification that has to be understood in deontological terms, or by arguing that even if *some* genuine notion of epistemic justification is captured by deontological accounts of justification, still there are other equally legitimate notions of justification that are not to be understood in deontological terms. Now, most epistemologists nowadays endorse non-deontological accounts of justification. Moreover, the monist assumption according to which there is a single correct account of epistemic justification has recently been vigorously challenged, both by internalists and externalists.[8] This is why, as it stands, the argument is unlikely to attract serious consideration from contemporary philosophers.

One of the chief merits of Lockie's discussion of IETAF is that it attempts to defend it without relying on MED. Lockie actually rejects MED and endorses a pluralist stance in epistemology according to which there is a

[6] I won't discuss this principle here. For a defence, see Smithies (2012).
[7] Lockie (2018) discusses many of them in Chapter 8.
[8] See Coliva and Pedersen (2017), especially the Introduction and the literature referred to therein.

plurality of accounts of epistemic justification that capture equally important dimensions of epistemic evaluation.[9] In particular, he makes room for an internalist deontological conception of justification, and an externalist (non-deontological) conception of justification. Lockie's strategy consists in showing that even if the determinist's belief is justified according to an externalist notion of justification, the determinist can't occupy a coherent theoretical stance unless the belief in determinism is also justified$_{ED}$. Lockie's remarks on the conclusion of his argument give us a hint that indicates the missing premise that puts pressure on the determinist:

> Determinists must be able to justify their position and oppose their opponents' positions. The framework for such justification must be in place – no metaphysics can be so powerful, so totalizing, as to undermine it. (Lockie 2018, 183)

In claiming that "[d]eterminists must be able to justify their position and oppose their opponents' positions", Lockie seems to suggest that this justificatory ability involves the appeal to epistemic obligations. To a first approximation, by holding the determinist view, the determinist is willy nilly committed to the claim that the opponent ought to abandon her own view and endorse determinism. If a contention along these lines is correct, then the determinist can't be content with a non-deontological (be that internalist or externalist) justification for her belief, for it is part of what it takes to justify one's own position and to oppose the opponent's position to hold that there are epistemic obligations of the sort posited by deontological accounts of justification. Thus, this is the crucial premise that Lockie needs in order to use IETAF to conclude that determinism is self-defeating:

Commitment$_{EO}$: in judging determinism to be true, the determinist is committed to take it that there are epistemic obligations.

Since a deontological conception of epistemic justification understands justification in terms of the satisfaction of epistemic obligations, a commitment to epistemic obligation is a commitment to the possibility of justified$_{ED}$ beliefs. If we add Commitment$_{EO}$ to premises P1-P5, we are then in a position to understand how IETAF is meant to yield the conclusion that determinism is self-defeating. In judging that determinism is true, the determinist is committed to take it that there are no epistemic obligations and thus no justified$_{ED}$ beliefs (via P1-P5); and yet she is at the same time committed to take it that there are epistemic obligations

[9] See Lockie (2018, chap. 2).

(Commitment$_{EO}$), and in particular that her own judgment in determinism fulfils one such obligation and thus counts as justified$_{ED}$ (via Commitment$_{EJ}$).

It is clear from Lockie's discussion of IETAF that IA is supposed to provide a motivation for Commitment$_{EO}$.[10] In what follows I shall argue that IA doesn't provide any motivation for Commitment$_{EO}$ but rather relies on it. I shall also distinguish and reject three different interpretations of Commitment$_{EO}$. On this ground, I will conclude that IETAF fails to show that determinism is self-defeating. Before coming to the critical evaluation IA and Commitment$_{EO}$, I shall discuss Lockie's own preferred version of epistemic deontologism, as this will play a crucial role in the evaluation of IA.

3. Lockie's Epistemic Deontologism

There is one well-known objection against epistemic deontologism. According to Alston (1988), we have epistemic obligations to believe only if we have direct voluntary control over the formation of our beliefs, but since we lack this control we have no such epistemic obligations.

Lockie himself endorses doxastic involuntarism, i.e., the claim that we do not possess direct voluntary control over the formation of our beliefs. However, he addresses Alston's objection by making two moves: by shifting the focus of epistemic obligations from belief to the whole process of inquiry that culminates with the (involuntary) formation of belief; and by arguing that we do possess the kind of freedom that is required to underpin these obligations.

Lockie doesn't offer a detailed account of our epistemic obligations, but we can appreciate what he thinks about the issue by considering the following paradigmatic case of deontological appraisal:

> Envisage a detective who has, throughout his police career, demonstrated a poor attitude, being lazy, egotistical, lacking due diligence, lacking moral seriousness and possessing a laissez-faire approach to his professional duties. […] Through assiduous flattery and unctuous professional networking, our detective becomes lead investigator in a murder investigation, where he fails to seal the crime scene early enough, he cross-

[10] See Lockie's first option among the moves that are available to the determinist to reject the argument. The move consists in showing that "epistemic normativity is not to be understood on the model of 'oughts'" (Lockie 2018, 184).

> contaminates the storage of DNA evidence and he fails to systematically track down, cross-reference and record the relevant witness statements. He also fails to study the witness statements and forensic evidence with sufficient rigour and intricacy, or think carefully and systematically enough about the evidence and the unfolding investigation in the way he has been trained throughout his police career. He believes what he subsequently believes ('suspect x did it!') with sincere conviction – but he is unjustified (deontically) because of his deplorable cognitive conduct, his wholesale epistemic irresponsibility. Let us suppose his late-stage final processes of belief formation and fixation (say, the micro-cognitive processes that occur subsequent to his poor conduct, intellectual or otherwise) are entirely involuntary; still, he is epistemically unjustified in a strongly deontic sense. (Lockie 2018, 47-8)

With this case in mind, we can distinguish between two kinds of epistemic obligations and clarify the scope of Lockie's view.

Doxastic obligations are those obligations that concern *what* subjects ought to believe – as in the case in which one ought to judge that *p*, say, where *p* follows from truths believed by the subject and the subject knows that *p* follows from them.[11] According to Alston's objection, these obligations exist only if we possess the kind of voluntary control over the formation of belief which is denied by doxastic involuntarism. Since Lockie is a doxastic involuntarist, Lockie's deontologism is bound to reject the existence of doxastic obligations.

In fact, the detective case does not feature doxastic obligations, but rather what we might call *cognitive obligations*, that is, obligations that concern the way in which subjects ought to conduct their cognitive activities for the sake of inquiry – as in the case of the obligation to be systematic and careful in one's search for evidence, say.[12] Although we don't have the freedom that is required to underpin doxastic obligations, Lockie argues

[11] The choice of this principle is just for illustrative purposes. I am not concerned here with the content of specific epistemic obligations, but with the general contention that there are epistemic obligations at all and that these are presupposed in the activity of judging.

[12] These obligations are sometimes described in the literature as 'intellectual obligations'. See Alston (1988) for a characterization of intellectual obligations. For more on this topic and the varieties of epistemic deontologism, see Vahid (1998), Nottelmann (2013), and Peels (2017). In this paper, I prefer to distinguish between doxastic and cognitive obligations in order to leave it an open question whether Lockie's favoured obligations are what Alston and others have described as intellectual obligations.

that we do have the freedom that is required in order to fulfil cognitive obligations of the sort described in the detective case.

With these clarifications in mind, we can turn to IA.

4. Deontological Appraisal as Ineliminable

As I understand IA, its aim is to establish the following thesis:

Ineliminability deontological appraisal [IDA]: We are bound to presuppose that we have some epistemic obligations.[13]

According to IDA, deontological appraisal is ineliminable not so much because there *are* some epistemic obligations; rather, deontological appraisal is ineliminable because we are bound to *presuppose* that there are some epistemic obligations.

Lockie considers two ways in which one can argue for the eliminability of deontological appraisal. One is to hold that our "epistemic obligations may be so limited as to be uninteresting".[14] Deontological appraisal would be eliminable on that view because its scope of application would so limited as to be uninteresting. This is the weaker challenge to his view, and Lockie has two responses to it which do not require the appeal to IA.[15]

According to the second way of eliminating epistemic deontologism "the entire framework of [deontological] internalist justification is abandoned for the entire framework of externalist epistemic value".[16] This is the kind of challenge to which IA is supposed to provide an answer. This challenge can in turn be understood in at least two relevant ways:

---

[13] Although I present IA as an argument for the ineliminability of *deontological appraisal*, Lockie presents it as an argument for the ineliminability of *internalism*, or as a transcendental argument against a *totalizing externalism.* However, Lockie makes clear in several occasions (e.g., Lockie 2018, 28) that by 'internalism' he refers to the internalist deontological conception of justification. Moreover, Lockie also says (e.g., Lockie 2018, 118) that he prefers to deploy his argument in connection with obligations in particular, rather than in connection with internalism in general, although he eventually also presents an argument for the ineliminability of a non-deontological form of access internalism.

[14] Lockie (2018, 115).

[15] The first response is that we do in fact possess a significant amount of control over our cognition (Lockie 2018, chap. 4). The second response is that "However limited our agency and access may seem when considered from without, considered from within an epistemic perspective these are all the resources we have; and any limitations of these resources will leave unaffected the importance of doing the best we can" (Lockie 2018, 117)

[16] Lockie (2018, 116).

1. Deontological appraisal might be abandoned because one discovers that epistemology has nothing to do with deontology, or that justification is not to be understood in deontological terms. One way of pressing this objection against Lockie is by arguing that cognitive obligations are not genuinely epistemic, or that they have nothing to do with epistemic justification.[17]

2. Another way of reading the claim about the complete elimination of deontological appraisal is to take it as the claim that we lack the kind of control that is required to underpin our supposed epistemic obligations. This challenge is precisely the one that the determinist is raising: by arguing for determinism, the determinist is in a position to argue that there are no epistemic obligations.

With these clarifications in mind, we can better appreciate the nature of IA. Its aim is to show that *even if* (1) and (2) are correct, still we are bound to proceed *as if* we had epistemic obligations. So, to illustrate with (2), which is the central case in the context of a transcendental argument against the determinist, even if it is true that we lack the freedom needed to underpin epistemic obligations, still we are bound to presuppose that we have some epistemic obligations.

5. The Transcendental Argument for the Ineliminability of Deontological Appraisal

What is the argument for IDA? Lockie first provides an argument for the ineliminability of a non-deontological form of access internalism and then extends this argument to deontological appraisal. The central insight of the argument is expressed by Lockie as follows:

> However limited psychological science shows us to be, we cannot be so limited as to undermine the ability of such scientists to uncover our limits, then recommend (pessimistic) conclusions for epistemology based on such discoveries. On the assumption that they must have access to the ground for maintaining how limited we are in our access, there must be a limit to those limits. (Lockie 2018, 118)

[17] This is a standard challenge to epistemic deontologist views that do not focus on doxastic obligations but on intellectual or cognitive obligations. See Alston (1988, sec. VII) for a version of the challenge. See Peels (2017) for an answer to Alston's objection. See Lockie (2018, chap. 3) for his answer to this challenge.

This passage suggests that there is at least one dimension of epistemic evaluation that is ineliminable, as it would always be possible to evaluate the epistemic credentials of our beliefs by checking the quality of our grounds for them. This internalist dimension of evaluation can't be coherently rejected: however limited we end up to be, there must be some *ground* on the basis of which the objector claims that we are so limited, and thus the objector's belief itself can be evaluated by checking whether her grounds are epistemically good enough.

After presenting this argument, Lockie states that "[w]hat goes for access and control, goes for obligation".[18] His argument here is very compressed, but its crucial insight is captured by the following observation:

> It is indefensible to suppose we could abandon the last epistemic 'ought' for a wholly externalist conception of epistemic value, as the last 'ought' is the ought that urges us to eliminate itself. (Lockie 2018, 119)

Lockie doesn't provide further explanations of the nature of the claim that is made here, so one is left with several questions: What is exactly the "last 'ought'"? And what does it mean that "the last 'ought' is the ought that urges us to eliminate itself?".

In what follows I shall read Lockie's point in the last quoted passage as expressing the endorsement of $\text{Commitment}_{\text{EO}}$. According to this reading of the argument, Lockie is suggesting that in arguing for the abandonment of an ought-based epistemology one is thereby committing herself to the existence of epistemic obligations. So, coming back to IETAF, according to this reading of IA the determinist is someone who is arguing for the abandonment of an ought-based epistemology, and by so arguing she is committed to the existence of epistemic obligations.

6. No Commitment to Epistemic Obligations

In order to assess whether $\text{Commitment}_{\text{EO}}$ is true I shall rely on the following quite liberal understanding of how judgment's commitments work. A subject's judgment that *p* is committed to the truth of some proposition *q* (if and) only if it is not rational (or possible) for the subject to judge that *p* while she is at the same time judging that *q* is false or while she is at the same time open-minded as to whether *q* is true or not. We can then test a candidate judgment's commitment to judge that *q* by

[18] Lockie (2018, 118).

asking whether it would be rational (or possible) for the subject to judge that *p* while also judging that *q* is false (or while also being open-minded as to whether *q* is true). If it is rational (or possible) to judge that *q* is false (or to be open-minded as to whether *q* is true) while judging that *p*, then in judging that *p* one is not thereby committed to judge that *q*. On the other hand, if it is not rational (or possible) to judge that *q* is false (or to be open-minded as to whether *q* is true) while judging that *p*, then we have (arguably conclusive) grounds to conclude that in judging that *p* we are committed to judge that *q*.

To illustrate, it would not be rational (or even possible) for a subject to judge that *p* while at the same time judging that there are no evidential grounds whatsoever for *p*. There is something Moore-paradoxical in judging that *p* and that there are no evidential grounds for *p*. For, if there are no evidential grounds for *p*, then from the subject's first personal point of view it is entirely arbitrary to regard *p* as true (as opposed to any other proposition incompatible with *p*). This provides evidence for taking it that in judging that *p* one is thereby committed to take it that there are evidential grounds for *p*.[19]

With this understanding of judgment's commitments in mind, we can test the various interpretations of $\text{Commitment}_{\text{EO}}$. We get two versions of $\text{Commitment}_{\text{EO}}$ by distinguishing between doxastic and cognitive obligations:

$\text{Commitment}_{\text{DO}}$: In judging determinism to be true, the determinist is thereby committed to take it that there are doxastic obligations.

$\text{Commitment}_{\text{CO}}$: In judging determinism to be true, the determinist is thereby committed to take it that there are cognitive obligations.

$\text{Commitment}_{\text{CO}}$ seems to fail the commitment test. To appreciate the point, contrast these cases: (a) the detective claims that *p*, and then is asked whether there is any evidential ground for taking *p* to be true; (b) the detective claims that *p* and then is asked whether he *has been* diligent and systematic in his inquiry; (c) the detective claims that *p*, and then is asked whether he *ought to be* diligent in his inquiry (or whether he is under any of the many cognitive obligations that Lockie considers in his detective case). In case (a), as we have just seen, it is clear that it would not be rational (or even possible) for the detective to judge that *p* while judging that he has no evidential grounds for *p* (or while being open minded about

[19] Compare with Smithies (2012) who proposes a similar argument that appeals to Moore-paradoxicality and a similar account of how commitments work.

whether he possesses evidential grounds for *p*). In case (b), it would also seem not to be rational, for the detective, to judge that he was not diligent and systematic in his inquiry, since by so judging he would thereby be in a position to doubt whether he genuinely possesses good evidential grounds for *p*. If he had not been diligent and systematic, he might have missed some fundamental piece of evidence, or he might have misunderstood the available evidence. And if this is the case, his very judgment that *p* is jeopardized, as the detective is in a position to doubt whether his grounds for judging that *p* are good enough. However, and this is the crucial point, it is one thing for the detective to judge that he was diligent and systematic, and it is another thing for him to judge that he *ought* to be diligent and systematic. These are two separate issues: a subject might be diligent and systematic in one's inquiry, and she might end up in a position in which she judges that *p* on the basis of good evidential grounds, and yet she can at the same time deny, for reasons like those proposed by the determinist, say, that there is an obligation to be diligent and systematic. Thus, it is entirely possible and rational for the detective to judge that *p*, that he possesses good evidential grounds for *p*, that he was diligent and systematic in his inquiry, while also denying that he ought to be diligent and systematic in his inquiry. In judging that *p* we do not seem to commit ourselves to the existence of cognitive obligations. $\text{Commitment}_{\text{CO}}$ fails the commitment test.

When compared to $\text{Commitment}_{\text{CO}}$, $\text{Commitment}_{\text{DO}}$ seems to enjoy some prima facie plausibility. The intuitive ground for $\text{Commitment}_{\text{DO}}$ is that in judging that *p* we seem to be *recommending p* as the content to be judged, and this might be captured by saying that in judging that *p* we are committed to the existence of a doxastic obligation to the effect that one ought to judge that *p*, or something along these lines.

First of all, even if we concede, for the argument's sake, the prima facie plausibility of $\text{Commitment}_{\text{DO}}$, Lockie can't avail himself of this move. For, Lockie endorses doxastic involuntarism, and thus he grants that we lack the freedom needed to underpin these doxastic oughts. Therefore, if Lockie were suggesting to rely on $\text{Commitment}_{\text{DO}}$ for his IETAF, he would end up occupying the same self-defeating position that he is attributing to the determinist:[20] he holds doxastic involuntarism, and yet by holding it he is committed to doxastic voluntarism, as he is committed to the existence of doxastic obligations which require the sort of freedom posited by doxastic voluntarism.

[20] I will come back to this problem in §9 by referring it to the overall transcendental argumentative strategy employed by Lockie.

This is a problem which relates to Lockie's overall view, namely his acceptance of doxastic involuntarism and his consequent rejection of doxastic obligations. One might wish to endorse this reading of IA either by also arguing for doxastic voluntarism, or by denying that there is anything problematic in holding the self-defeating stance which Lockie occupies.

Be that as it may, $\text{Commitment}_{\text{DO}}$ also fails the commitment test. A subject might be rational in holding that *p* while comprehendingly denying that she ought to judge that *p*. Consider case (c) again. In this case, the detective might judge that he was diligent and systematic in his inquiry, and he might also judge that he possesses good evidential grounds for judging that *p*. However, all of this is compatible with the fact that there is no obligation to judge that *p*. One might reject the existence of such obligation by endorsing the impossibility of judging otherwise which is required in order for there to be obligations at all (at least in so far as obligations are understood within an incompatibilist framework, which is the only one pertinent here).[21]

Moreover, and relatedly, the intuitive ground for $\text{Commitment}_{\text{DO}}$ might be explained by appealing to normative commitments that do not involve obligations. Consider the following:

Judgment's Commitment to Alethic Correctness ($\text{Commitment}_{\text{AC}}$): in judging that *p* one is thereby committed to take one's judgment that *p* as correct.

Judgment's Commitment to Good Evidential Grounds ($\text{Commitment}_{\text{GEG}}$): in judging that *p* one is thereby committed to take it that she possesses good evidential grounds for *p.*

I have argued before that $\text{Commitment}_{\text{GEG}}$ is true.[22] $\text{Commitment}_{\text{AC}}$ is also arguably true, as it is reasonable to suppose that competent believers are sensitive to the truth of the following principle about the normative connection between truth and judgment:

[21] Notice that judging that it is not the case that one ought to judge that *p* does not entail that one ought not to judge that *p*. Judging that *p* and that one ought not to judge that *p* is indeed Moore-paradoxical, but judging that *p* and that it is not the case that one ought to judge that *p* is not.

[22] See also Smithies (2012) for a similar argument.

Alethic Correctness: a judgment that *p* is correct (if and) only if *p* is true.[23]

Given Commitment_{AC} and Commitment_{GEG} we might capture the intuitive thought according to which in judging that determinism is true the determinist is also somehow recommending determinism as the view to be believed. By judging that determinism is true the determinist is committed to the possession of good evidential grounds for so judging.[24] This commitment captures the sense in which a determinist is inviting the opponent to agree with her, as she is claiming to be judging a proposition that is well supported by the evidence, and thus she is claiming that her belief is *justified*, or *rational*. Moreover, the determinist is also committed to take it that it is *correct* to judge as she does. Thus, in this sense, the determinist is recommending the opponent to be a determinist, since to judge in determinism is the correct attitude to have with respect to the issue whether determinism is true or not. Crucially, none of these normative commitments and none of the normative notions they involve (correctness, justification, rationality) require the existence of epistemic obligations.

Although I do not claim to have provided a full vindication of Commitment_{AC} and Commitment_{GEG}, I think that since it is available to the determinist to appeal to them in explaining her normative commitments, a defender of IETAF must provide arguments to show that Commitment_{DO} is true and can't take it for granted in her argument against the determinist.

7. The Last Duty

I have understood epistemic obligations as specific obligations concerning what ought to be believed or how one ought to conduct one's own inquiry, and I have asked whether in judging determinism to be true the determinist is committed to any such *specific* epistemic obligation. However, there is another way of reading Lockie's IA.

Lockie argues that there is a *single* overarching obligation from which all other more specific obligations follow.

[23] See Wright (1992) and Lynch (2009) who take this principle to be an a priori platitude or truism about truth. See Ferrari (2018) and the literature referred to therein about the variety of understanding of the truth-norm for judgment.

[24] This is also the conclusion of Lockie's transcendental argument for the ineliminability of a non-deontological conception of internalism that I have summarized above.

> The only fundamental internalist 'ought' is (early) Chisholm's 'primary intellectual duty' to aim to acquire truth and avoid error. That is as much content to the notion of duty as internalism as such need make space for. Given such an approach to epistemic duty, it becomes optional whether the proponents of any particular internalist account wish to articulate, at the level of first-order epistemic theory, any system of duties, rules, etc. Given that one ought to aim to possess truth, developing an account of the means to that end then becomes an engineering problem. (Lockie 2018, 111)

Crucially, the only fundamental epistemic duty is not to actually possess the truth,[25] but rather to *aim* to possess it. On this account, being epistemically justified is then a matter of doing the best that one can in order to fulfil the overall obligation to aim to believe the truth. According to this view, there is no need to specify further independent epistemic obligations beside the fundamental epistemic duty to aim to believe the truth. Arguably, the specific duties that we have in specific cases can all be derived from the last duty by asking what ought to be done in order to aim to believe the truth – and this is an "engineering problem".[26]

With this account in mind, we can now re-read the whole passage in which Lockie argues that this overarching duty is ineliminable:

> It was stressed above that internalism should be understood as a very high-order theory, not the claim that we must be operating on a set of first-order rules or obligations … So, the crucial question is this: at the very high-order level, is the last, most fundamental epistemic 'ought' ineliminable, foundational, sui generis, or is not? We have to hold that it is ineliminable. It is indefensible to suppose we could abandon the last epistemic 'ought' for a wholly externalist conception of epistemic value, as the last 'ought' is the ought that urges us to eliminate itself. (Lockie 2018, 118-9)

According to the present reading of the argument, the "last most fundamental epistemic 'ought'" is the duty to aim to achieve the truth. To abandon the last duty consists in judging that our supposed last duty is not a duty at all – and thus that it is false that we ought to aim to believe the

[25] For the sake of simplicity, I will speak of the aim to possess/achieve/believe the truth only, and will drop the talk of 'truth maximization' and 'error avoidance', since they are not crucial in this context and their omission won't affect the argument. See Lockie (2018, 5.1.2) for more on this point.
[26] Lockie (2018, 111-112).

truth. Now, with this interpretation in mind, the argument seems to amount to the following: in judging that we are not under the obligation to aim to believe the truth one is thereby implicitly presupposing that we are under the obligation to aim to believe the truth. For, it is in the name of the duty to aim to believe the truth that one claims that (it is *true* that) there is no such thing as the duty to aim to achieve the truth. It is in aiming to judge the truth about deontology that one eventually ends up judging that there is no such thing as the obligation to aim to judge truly. So understood, the argument relies on the following specification of $\text{Commitment}_{\text{EO}}$:

Determinist's Commitment to Last Duty ($\text{Commitment}_{\text{LD}}$): in judging determinism to be true, the determinist is committing herself to the existence of the last duty, namely to it being the case that we ought to aim at achieving the truth.

Is there any prima facie plausibility in the claim that in judging we are committed to take it that we ought to aim to achieve the truth?[27] One might wish to argue for $\text{Commitment}_{\text{LD}}$ by noticing that to judge that *p* is to take a commitment towards the *truth* of *p*. Moreover, some theorists argue that judging aims at truth,[28] and on this ground one might argue that since judging aims at truth, by issuing a specific judgment (like the judgment in determinism) one is thereby committed to the existence of the corresponding obligation to aim at truth. However, to aim at X does not need to generate a commitment to an *obligation* to aim at X. Compare with archery. By aiming at doing center I am not thereby committed to take it that one has an obligation to aim at doing center. It is entirely possible to aim at doing center while consciously rejecting any obligation to aim at doing center. One might insist that there is a conditional obligation there: in so far as you want or have a reason to do archery, you ought to aim at doing center. However, one can consistently aim at doing center by taking it that doing center is *correct*, or *good* (at least in archery), and yet deny the existence of obligations (on determinist grounds, say). Analogously, it is rational for a subject to aim to believe the truth about free will, to eventually conclude that determinism is true, and to reject, on this ground, the existence of an obligation to aim at truth, while at the same time conceding that her judgment in determinism is correct ($\text{Commitment}_{\text{AC}}$), that it is well grounded ($\text{Commitment}_{\text{GEG}}$), and that judging truly is good

[27] Lockie argues for the claim that the duty to aim at truth is the fundamental duty, but he doesn't provide any support for $\text{Commitment}_{\text{LD}}$. For the sake of the present argument, I will concede Lockie's claim about what our most fundamental duty is, and will concentrate the discussion on $\text{Commitment}_{\text{LD}}$ only.

[28] Various understanding of this claim are argued for by Steglich-Petersen (2006), Bird (2007), Velleman (2000). For criticisms, see Shah (2003), Owens (2003), and Zalabardo (2010).

or valuable. As we have noticed above, the intuitive ground for the various readings of Commitment_{EO} might be captured by appealing to non-deontological normative commitments involving the notion of correctness and the non-deontological epistemic notions of justification and rationality.

A defender of Commitment_{LD} might understand the last duty as the duty to aim at truth *while inquiring,* and not as the claim that *in judging* we ought to aim at truth. Since the aim of inquiry is to believe truly, one might argue that in inquiry we ought to aim at achieving truth and one might further argue that we are committed to the existence of this obligation in inquiry. But then the claim suffers from the same problem that we noticed in the case of Commitment_{CO}. The last duty is meant to be the source for the sort of cognitive obligations that are appealed to in the detective case. We ought to be systematic in our search for evidence, say, because we ought to aim to achieve the truth, and being systematic is what it takes to aim to achieve the truth. But since it is entirely possible to rationally judge that *p* while denying the existence of specific cognitive obligations, it is also rational to judge that *p* while denying the source of specific cognitive obligations. So, I conclude, Commitment_{LD} is false.

8. IETAF and Modest Transcendental Arguments

Thus far, I have argued that Commitment_{EO} is false, and that as a result IETAF fails to deliver the conclusion that determinism is self-defeating. In concluding the paper, I wish to highlight another important limitation of Lockie's transcendental strategy which arises even if we concede that his IETAF succeeds.

Let us suppose, for the argument's sake, that IETAF succeeds in showing that determinism is self-defeating. However, IETAF is compatible with the fact that we have good grounds for believing that determinism is true. The resulting stance would be such that one is unavoidably committed to presuppose the truth of a proposition – the existence of epistemic obligations, and therefore the existence of (incompatibilist) freedom – even if one appreciates that all the evidence indicates the falsity of that proposition. Within this stance, we might keep being confident that determinism is true, even if we realise that by being so confident we are also presupposing that there is the sort of free will whose existence is denied by determinism.

In order to appreciate why this is an important limitation, it is useful to locate IETAF in the debate between modest and ambitious transcendental

arguments.[29] Ambitious transcendental arguments aim at showing that the *truth* of some proposition *p* is the condition of possibility for some fact that even the sceptical opponent is prepared to accept. Modest transcendental arguments aim at showing that *to believe in* (or have some *cognitive relation* towards) some proposition *p* is a condition of possibility for the fact which is agreed upon by sceptics and non-sceptics alike. IETAF belongs to the category of modest transcendental arguments since its aim is to show that to *presuppose* the existence of free will is something that we do whenever we judge, and in this sense is a condition of possibility of the very activity of judging.

Now, there is an important disanalogy between Lockie's dialectical engagement with the determinist and canonical uses of the modest transcendental strategy. Modest transcendental arguments have been often explored as viable strategies to respond to a *sceptic* who is challenging the possibility of knowing some proposition *p*, and not as strategies to respond to someone *denying* the truth of some proposition *p.* To illustrate, modest transcendental arguments have often been used in order to respond to the sceptic about the existence of the external world.[30] Crucially, this sceptic claims that we do not have enough grounds (or grounds at all) to believe that the external world exists, but he does not argue that we have good grounds to believe that the external world doesn't exist. This is why a modest strategy is (modestly) satisfying in this context: because a modest transcendental arguer ends up in a position in which she does not have grounds for believing *nor does she have grounds for disbelieving* in the existence of the external world, and yet she is bound to believe in its existence as this belief being in place is a condition of possibility for some inescapable cognitive activities (like judging, experiencing, etc.) whose reality is conceded by the sceptic herself.[31] But the fight against the determinist is different. To continue the comparison, the determinist is like an idealist denying the existence of the external world. The determinist is not claiming that we do not have enough reasons to settle the question whether there is free will or not. The determinist is claiming that free will doesn't exist, and she takes herself to have good

[29] See Stroud (2000), Stern (2017) and the literature referred to therein.

[30] See Stern (2000) and those hinge epistemologists like Strawson (1985), Wright (2004), and Coliva (2015) who appeal to Wittgenstein's remarks in *On certainty* in order to answer to external world scepticism.

[31] Moreover, some modest transcendental arguers go further. Stern (2000) argues that the belief in the external world is warranted; Wright (2004) and Coliva (2015) claim that belief in the external world enjoys a special kind of non-evidential warrant; Pritchard (2016) claims that the proposition that there is an external world is beyond the scope of rational evaluation. No such claims are made by the indirect transcendental arguer for freedom. This further reinforces the point I am making here about the disanalogy between IETAF and modest transcendental arguments against the external world sceptic.

grounds for that claim. So, the indirect transcendental arguer for freedom will end up endorsing a deeply dissatisfying standpoint:[32] even if we can't but presuppose that we possess free will, we have very good reasons – as the determinist says – to believe that this unavoidable presupposition is in fact false. This standpoint is in no way intellectually reassuring: it rather represents our cognitive standpoint like a cage which is structured by false unavoidable presuppositions. This might be the truth about our condition – although I have offered reasons to think that it is not. But this is in no way a truth that allows us to claim victory over the denier of free will.

Acknowledgements

An earlier draft of this paper has been presented in Bologna, in 2019, at COGITO Research Center. I am grateful to all people in attendance for helpful comments and suggestions. Special thanks are due to Matti Eklund, Filippo Ferrari, Sebastiano Moruzzi, Elena Tassoni, and Giorgio Volpe.

REFERENCES

Alston, W. P. 1988. The deontological conception of epistemic justification. *Philosophical Perspectives* 2: 257-299.

Alston, W. P. 2005. *Beyond Justification: Dimensions of Epistemic Evaluation*. Ithaca: Cornell University Press.

Bird, A. 2007. Justified judging. *Philosophy and Phenomenological Research* 74: 81-110.

Boyle, J., G. Grisez, and O. Tollefsen. 1976. *Free Choice: A Self-Referential Argument*. Notre Dame: University of Notre Dame Press.

Coliva, A. 2015. *Extended Rationality: A Hinge Epistemology*. Basingstoke: Palgrave-Macmillan.

Coliva, A., and N. J. L. L. Pedersen, eds. 2017. *Epistemic Pluralism*. Basingstoke: Palgrave-Macmillan.

Epicurus 1926. *The Extant Remains*, ed. C. Bailey. Oxford: Clarendon Press.

Ferrari, F. 2018. Normative Alethic Pluralism. In *Pluralisms: Truth and Logic,* eds. N. Kellen, N. J. L. L. Pedersen, and J. Wyatt, 145-168. London, UK: Palgrave Macmillan.

[32] Unless, of course, the modest transcendental strategy is coupled with arguments that show that we have no grounds at all in favour of determinism. But this would be an altogether different project.

Honderich, T. 1990a. *A Theory of Determinism, vol. 1: Mind and Brain*. Oxford: Clarendon Press.

Honderich, T. 1990b. *A Theory of Determinism, vol. 2: The Consequences of Determinism.* Oxford: Clarendon Press.

Jordan, J. 1969. Determinism's dilemma. *The Review of Metaphysics* 23/1: 48-66.

Knaster, S. M. 1986. How the self-defeating argument against determinism defeats itself. *Dialogue* 25: 239-244.

Lynch, M. P. 2009. *Truth as One and Many*. Oxford: Clarendon Press.

Nottelmann, N. 2013. The deontological conception of epistemic justification: A reassessment. *Synthese* 190: 2219-2241.

Owens, D. J. 2003. Does belief have an aim? *Philosophical Studies* 115: 283-305.

Peels, R. 2017. Responsible belief and epistemic justification. *Synthese* 194: 2895-2915.

Pritchard, D. 2016. *Epistemic Angst: Radical Skepticism and the Groundless of Our Believing*. Princeton: Princeton University Press.

Lockie, R. 2018. *Free Will and Epistemology. A Defence of the Transcendental Argument for Freedom.* London, Oxford, and New York: Bloomsbury Academic.

Shah, N. 2003. How truth governs belief. *Philosophical Review* 112: 447–482.

Shah, N., and D. Velleman. 2005. Doxastic deliberation. *Philosophical Review* 114: 497–534.

Slagle, J. 2016. *The Epistemological Skyhook: Determinism, Naturalism, and Self-Defeat*. New York: Routledge.

Smithies, D. 2012. Moore's paradox and the accessibility of justification. *Philosophy and Phenomenological Research* 85: 273-300.

Steglich-Petersen, A. 2006. No norm needed: On the aim of belief. *Philosophical Quarterly* 56: 499–516.

Stern, R. ed. 1999. *Transcendental Arguments: Problems and Prospects*. Oxford: Oxford University Press.

Stern, R. 2000. *Transcendental Arguments and Scepticism: Answering the Question of Justification*. Oxford: Oxford University Press.

Stern, R. 2017. Transcendental arguments. *The Stanford Encyclopedia of Philosophy (Summer 2017 Edition)*, Edward N. Zalta (ed.), Accessed May 30, 2019. https://plato.stanford.edu/archives/sum2017/entries/transcendental-arguments/

Strawson, P. F. 1985. *Scepticism and Naturalism: Some Varieties*. London: Routledge.

Stroud, B. 1968. Transcendental arguments. *Journal of Philosophy* 65: 241-256. Reprinted in Stroud 2000.

Stroud, B. 2000. The goal of transcendental arguments. In his *Understanding Human Knowledge: Philosophical Essays*, 203–223. Oxford: Oxford University Press.

Vahid, H. 1998. Deontic vs. nondeontic conceptions of epistemic justification. *Erkenntnis* 49: 285-301.

Velleman, D. 2000. On the aim of belief. In his *The Possibility of Practical reason*, 244–281. New York: Oxford University Press.

Wedgwood, R. 2002. The aim of belief. *Philosophical Perspectives* 36 (s16): 267-297.

Wittgenstein, L. 1969. *On Certainty.* Oxford: Oxford Basil Blackwell.

Wright, C. 1992. *Truth and Objectivity*. Cambridge, MA: Harvard University Press.

Wright, C. 2004. Warrant for nothing (and foundations for free)? *Aristotelian Society Supplementary* 78: 167-212.

Zalabardo, J. L. 2010. Why believe the truth? Shah and Velleman on the aim of belief. *Philosophical Explorations* 13: 1–21.

Zanetti, L. 2019. Review: Robert Lockie, Free will and epistemology: A defence of the transcendental argument for freedom. *Dialectica* 73: 273-279.

EuJAP | Vol. 15 | No. 2 | 2019
UDK: 165.72 123:17
https://doi.org/10.31820/ejap.15.2.3

IS FREE WILL SCEPTICISM SELF-DEFEATING?

Simon-Pierre Chevarie-Cossette
King's College London

Original scientific article – Received: 29/05/2019 Accepted: 11/11/2019

ABSTRACT

Free will sceptics deny the existence of free will, that is the command or control necessary for moral responsibility. Epicureans allege that this denial is somehow self-defeating. To interpret the Epicurean allegation charitably, we must first realise that it is propositional attitudes like beliefs and not propositions themselves which can be self-defeating. So, believing in free will scepticism might be self-defeating. The charge becomes more plausible because, as Epicurus insightfully recognised, there is a strong connection between conduct and belief—and so between the content of free will scepticism (since it is about conduct) and the attitude of believing it. Second, we must realise that an attitude can be self-defeating relative to certain grounds. This means that it might be self-defeating to be a free will sceptic on certain grounds, such as the putative fact that we lack leeway or sourcehood. This charge is much more interesting because of the epistemic importance of leeway and sourcehood. Ultimately, the Epicurean charge of self-defeat fails. Yet, it delivers important lessons to the sceptic. The most important of them is that free will sceptics should either accept the existence of leeway or reject the principle that '"ought" implies "can"'.

Keywords*: Free will scepticism, self-defeat, self-refutation, leeway, sourcehood, Epicurus, "ought" implies "can", responsibility, reasons*

1. Introduction

Free will scepticism is the doctrine that we do not have free will, i.e. the kind of command or control of our own conduct that we would need in order to be morally responsible for it.[1] The concept of free will allows for different more precise *conceptions*. It is sometimes understood as the ability to choose amongst real alternatives or, as philosophers say, leeway;[2] and it is sometimes understood as being the source of one's conduct.[3] Thus, sceptical arguments typically conclude that free will does not exist, on the grounds that we lack leeway or that we are not the source of our actions.

Denying the existence of free will broadly defined might seem unthinkable to us. Radical as it stands, free will scepticism *must be wrong*.[4] But if it is so radical, could it be turned against itself like some other sweeping philosophical doctrines? Several philosophers in history took up this challenge.[5] Their arguments do not share a common philosophical lexicon. Nor do they always target free will scepticism explicitly—they often target determinism. But if they are sound, it is self-defeating to believe in free will scepticism, whether it is true or false. Since Epicurus first marshalled these arguments, let us call them 'Epicurean arguments.'

This paper is slightly unusual in that it does not belong to a contained contemporary debate where it could make a very specific contribution. Few philosophers after the 70s took Epicurean arguments seriously. Those who did proposed complex reconstructions which relied on a vast array of controversial views about epistemology and free will.[6] Going back to the simple argument of Epicurus, distilling its general idea, and using it to regiment refined but simple Epicurean arguments is my first goal. My second goal is to assess how the sceptic might respond to each type of argument and to unearth what she might already learn. I ask the reader to judge this critical engagement as it is: a precocious step in a nascent literature.

Here is the game plan to meet my two goals. I start with Epicurus' own argument (§2) and discuss what it means to be self-defeating. I then move to reconstructing the argument in the most charitable way (§3). This

[1] See e.g. Strawson (1994), Waller (2011), and Pereboom (2001; 2014a).
[2] See e.g. Vihvelin (2013) and van Inwagen (2017).
[3] See e.g. Frankfurt (1971) and Fischer (1994).
[4] For a recent discussion of the parallels between a dogmatic response to free will scepticism and to scepticism, see Chevarie-Cossette (forthcoming).
[5] See e.g. Hinman (1979), Dworkin (2011), Slagle (2016), and Lockie (2018).
[6] See e.g. Lockie (2018) and Slagle (2016).

gives rise to arguments related to sourcehood (§4) and leeway (§5), which I examine in turn.

2. Epicurus' Argument

About the view that all comes to pass by necessity, Epicurus argued:

> This sort of account (λόγος) is self-refuting (τρέπω περικάτω), and can never prove that everything is of the kind called 'necessitated'; but he [the sceptic or the determinist] debates this very question on the assumption that his opponent is himself responsible (δι ἑαυτοῦ) for talking nonsense. (Epicurus, *On Nature*. XXV, 34; translation Annas, Taylor, and Sedley 1983, 19–23)

To extract the best possible argument from this passage, we need to answer two questions. First, who does the Epicurean criticise: the determinist or the free will sceptic? Second, what does the Epicurean criticise this target for: making a discourse, maintaining a belief, or posing an action? For simplicity, call these two things respectively the *target* and the *object* of the Epicurean arguments.

2.1. The Target and Object of Epicurus' Argument

The official target of Epicurus is the Stoic who endorses the doctrine of necessity. This doctrine approximates *physical determinism*, the thesis that the conjunction of all the physical states in a given time slice and the laws of nature determines the states in any other time slice. However, Epicurus' argument is mostly worrying to determinists who also believe that determinism undermines free will, namely hard determinists—as opposed to determinists who do not, namely soft determinists.

Why? Epicurus' argument, as we will soon discover, turns on the claim that rationality or reason requires free will (whether sourcehood or leeway). The soft determinist may simply respond to Epicurus that the apparent, though illusory, difficulty of reconciling determinism with rationality is entirely inherited from the apparent, though illusory, opposition between determinism and free will. But this response is unavailable to the hard determinist. The hard determinist must be an *incompatibilist* between determinism and free will but a *compatibilist* between determinism and rationality. This is the real challenge. So, there is no doubt that, if Epicurean arguments work, they work against hard determinists. The most charitable reconstruction of Epicurean arguments

thus reads them as primarily targeting hard determinists.[7] To simplify, I will consider that the target of Epicurus extends to all free will sceptics. For one thing, sceptics typically admit the possibility of determinism; for another, Epicurus' argument hinges on the alleged consequences of determinism for free will (or responsibility), not on determinism itself.

Moving to the object of the argument. In the passage, Epicurus focusses on the sceptic's *logos* (λόγος). But *logos* is famously ambiguous and suggests at least two possible translations. It could mean 'account', i.e. the sceptic's doctrine; or it could mean 'reasoning', i.e. the sceptic's inference or belief from reasons. What is wrong with this account or with this reasoning? They are described as *trepetai egkalein* ('τρέπω περικάτω'), a technical term which literally translates as *defeating upside down*. Translators talk about 'self-refutation' or 'self-defeat', but what does that mean?

2.2. Self-Defeat and Self-Refutation

To understand Epicurus' argument, we must distinguish self-refutation from self-defeat and their respective objects.[8] A proposition (or an account) is self-refuting, in one common sense of the term, just when it is contradictory because it applies or refers to itself; for instance, 'No universal proposition is true'. Similarly, an argument is self-refuting just when it is unsound because it applies or refers to itself; for instance, 'single-premise arguments are invalid; therefore, single-premise arguments are invalid'.[9]

Free will scepticism is *not* self-refuting in this sense, since it is neither contradictory nor self-referential. The arguments supporting free will scepticism are not self-refuting either: while they might be unsound, this is not because they apply or refer to themselves. This suggests that when Epicurus attacks the sceptic's *logos*, he does not mean her 'account'.

However, there might be *epistemic* problems with *believing* or *reasoning* (in the sense of reasoning to a conclusion) that free will scepticism is true. After all, this is what Epicurus' argument suggests: that the content of a doctrine interacts with the epistemic attitudes of its proponent in an unfortunate way.

This is where the notion of self-defeat is relevant; for its object is precisely attitudes like beliefs or inferences rather than abstracta like propositions or

[7] Cf. Slagle (2016, 17, 28, 201–202) and Lockie (2018).
[8] In this, I follow Slagle's distinction (2016, 41–43).
[9] This conforms to Mackie's account (1964); see also Page (1992, 423).

accounts. There are two kinds of case of self-defeating propositional attitudes like belief. First, imagine Tommy, an undergraduate philosophy student who, fortunately for our purpose and unfortunately for his peers, likes to embrace the most radical theses he encounters. Today he believes that he has no beliefs. The fact that he believes this proposition ensures[10] that it is false. Tommy's belief is 'self'-defeating in the sense that the *believing* (the attitude) defeats *what is believed* (the content)—see **Act defeats content** below. Second, imagine that Tommy acquires the belief that he would not acquire a single justified belief that day. Even assuming that this is true, Tommy's belief is self-defeating. The presumed truth of Tommy's belief ensures that he believes it inadequately, e.g. without justification, irrationally, unreasonably, etc. Tommy's belief is 'self'-defeating in the sense that the truth of *what is believed* (the content) precludes the *believing* (the attitude) from being adequate—see **Content defeats act** below.

Thus, self-defeat is a property which, when it applies to beliefs, plays on an act/object ambiguity. It is either the *attitude of believing* which ensures that the *belief content* is false or, alternatively, it is the alleged truth of the *belief content* which ensures that the *attitude of believing* is inadequate. In light of this, it is no surprise that the object of Epicurus' argument was uneasy to identify.

A last remark: a belief is based on some grounds and those grounds are sometimes relevant to whether it is self-defeating. Last week, Tommy came to think that he was hopeless at remembering events in the distant past. That's fine. But Tommy formed his belief on the grounds that, when he was eight, he forgot his best friend's name four times. Now, *this* is problematic. Tommy's belief could well be true; and this time, it could be adequate if it was differently grounded. But the presumed truth of the belief content ensures that the believing is inadequate, *grounded as it is*. Thus, whether a belief is self-defeating is relative to the grounds on which it is believed.[11]

The following definition captures our remarks. (The same applies to inferences.)

[10] One might read 'ensures that' as 'is a sufficient reason for', 'explains why', 'grounds', or 'makes'.

[11] Some might think that this remark stretches the notion of self-defeat since the grounds of a belief is something external to it and so cannot contribute to *self*-defeat. Yet what a belief contains is contentious: a belief in the same proposition but which has been rebased on other grounds is not obviously the same belief. When I say that a belief is self-defeating, I use this broader view of what a belief contains.

> **Self-defeating beliefs** $=_{df}$ Suppose that a subject *S* believes that *p* on the grounds that *q*. *Then, S*'s believing that *p* is self-defeating *if and only* if *either*:
> **(Act defeats content)** The fact that *S* believes that *p* on the grounds that *q* ensures that *p* is false; or,
> **(Content defeats act)** The presumed truth of *p* ensures that *S* believes that *p* on the grounds that *q* inadequately.[12]

What is an adequate belief? For the purpose of this discussion, I will assume that it is a belief which is rational or reasonable. For our purpose, I find it useful to understand rationality in believing as believing according to some rules of rationality and reasonableness in believing as believing on good grounds or for good reasons.

I also think that self-defeating beliefs are beliefs that we should dispose of, at least upon recognising that they are such. This is because they are beliefs which generally fail to be reasonable. Typically, a self-defeating belief is unreasonable in that it implies that it is unsupported by reasons: so either it is false or unsupported by reasons. Upon discovering this fact, we lose any reason to maintain this belief (see Chevarie-Cossette 2019).[13]

3. Reconstructing the Epicurean Argument

We now have some tools to provide a charitable interpretation of Epicurus' claim.[14] Free will scepticism is not self-refuting, for there is nothing contradictory or self-referential about the *thesis* that free will does not exist. However, free will scepticism is a thesis that concerns conduct, but which extends to attitudes. The suggestion is that there is an unfortunate interaction between the content of the thesis of free will scepticism and the attitude of believing or inferring. This is crucial: a key to all Epicurean arguments is a connection between conduct and beliefs.

So, free will scepticism might be self-defeating to *believe in*. But if this is true, this cannot be in virtue of the fact that to believe it makes it false (as in Act Defeats Content). What remains is the possibility that the presumed

[12] I have presented this account elsewhere (Chevarie-Cossette 2019).
[13] We should note however that some philosophers still insist that they can reasonably question the existence of reasons (Olson 2014, chap. 9; cf. Cuneo 2007, chap. 4). The question turns essentially on whether reasons are inherently *normative* and on whether we can, in distinguishing reasonable from unreasonable beliefs, use a non-normative notion like evidence.
[14] There is also a plausible moral interpretation of Epicurus' argument, but I leave it aside for our purpose.

truth of free will scepticism (perhaps combined with something else) ensures that one does not rationally or reasonably believe it (Content Defeats Act). Why would this be true? The most straightforward explanation is that rationality or reasonableness requires free will.[15]

This explanation sits well with another remark of Epicurus on the topic:

> The man who says that all things come to pass by necessity cannot criticise (ἐγκαλέω) one who denies that all things come to pass by necessity: for he admits that this too happens of necessity. (Epicurus, *Extant Remains*, Frag. XL; translation Bailey 1926, 112)

According to this point, the sceptic cannot legitimately criticise her opponent. This is because in trying to persuade her opponent, the sceptic is appealing to her opponent's rationality; but, again, rationality implies free will.[16] So, in order to convince someone of a doctrine, the free wil sceptic must presume its falsehood.

We need not focus on this dialectical problem. *If* it is self-defeating to rationally persuade someone to be a sceptic, it has everything to do with the *rational* character of persuasion and nothing to do with its *interpersonal* character (cf. Castagnoli 2007, 16). So we should ask again: is it true that rationality or reason implies free will?

It seems, after all, that I can rationally believe that 2+2=4 without having free will, i.e. the command or control necessary for moral responsibility. Here are two *pro tanto* reasons to strengthen this impression. First, there seems to be no necessary opposition between rational belief and constrained belief. Quite the opposite, constraint seems like a part of rationality, as Nozick (1981, 4–8) and James noted (1912, 168–169): to give in to *forceful* or *knockdown* arguments, or to be *forced to* a conclusion, is not irrational at all. Second, there seems to be no necessary opposition between rational action and unfree action (in the sense of an *inner freedom*). A heroin addict in rehabilitation acts rationally in taking the safe dose given by a doctor, regardless of whether he has the command or control to refuse to take it. If there is no opposition between rationality on the one hand and coercion or absence of inner freedom on the other hand,

[15] By 'free will', it is important to insist that I mean nothing more than the command or control necessary for moral responsibility. Otherwise, this would be an anachronistic interpretation of Epicurus. See Bobzien (2000) and Frede (2011).

[16] Again, there is also a moral interpretation of this argument according to which it is unfair to criticise someone for not being a free will sceptic.

it is unclear that there is an opposition between rationality and absence of free will.

We can however refine our explanation of why it is self-defeating to believe in free will scepticism. Epicurus targeted the free will sceptic's *logos* or *reasoning*. But we reason *from premises* and we believe *on certain grounds*. Now, the grounds or premises for which one is a free will sceptic includes or implies the proposition that humans lack leeway or sourcehood. The Epicurean can take advantage of these points. She does not need to argue that the existence of rationality or reason implies the existence of free will, understood generally. She can instead argue more narrowly that either leeway or sourcehood implies rationally. Thus, believing in free will scepticism on the grounds that we lack leeway or sourcehood is self-defeating.

4. Sourcehood

Epicurus himself was a sort of sourcehood theorist. In fact, he claimed that the sceptic debates the question of responsibility 'on the assumption that his opponent is himself responsible (δι ἑαυτοῦ) for talking nonsense.' The Greek for 'responsible' is δι ἑαυτοῦ (*di eautou*), which means 'because of himself'—as opposed to because of something external (see Bobzien 2000, 291–292). This is sourcehood. So, according to Epicurus the sceptic cannot argue that we are never the *source of our conduct* because this presumes that she is not the *source of her reasoning or belief*. This is supposed to be self-defeating.

What is it to be the source of one's conduct? The contemporary literature gives us two main answers. We are the source of our conduct when our actions and omissions stem from reason-responsive mechanisms (see e.g. Fischer 1994; Hurley 2003; Sartorio 2016). Alternatively, we are the source of our conduct when our actions and omissions are the product of some of *our* desires or *our* values—something to which we identify or which belongs to our 'real selves' (see, e.g., Frankfurt 1971; Shoemaker 2015). So, the first view emphasises reason while the second emphasises identification or ownership.[17]

There is a sceptical concern corresponding roughly to each answer. The first is that reasons are irrelevant to the explanation of our conduct: our conduct, including our reasons for action, can be entirely accounted for in

[17] These answers are not exclusive: for instance, Fischer endorses an endorsement condition on responsibility.

terms of physical events. Call this *the exclusion concern*. The second concern is that our conduct does not ultimately belong to our real selves—or if it does, this is partly insignificant—because our conduct is ultimately determined by events outside of our control. The same goes for reasons for which we act. Call this the *ownership concern.* The two concerns are importantly linked: the exclusion concern makes the radical claim that we never act *for* reasons; the ownership concern leads to the less radical view that we never act for reasons that are ultimately *ours*. In a word, the first concern casts doubt on reason for conduct; the second on ownership of these reasons. Each sceptical concern gives rise to an Epicurean argument.

4.1. Exclusion

The exclusion concern about reasons stems from so-called *exclusion arguments* about mental states (see, e.g., Kim 2007). Roughly, mental states seem to be unidentical to physical states and yet realised in them. But then the cause of our conduct could be entirely accounted for in terms of physical states. This leaves no causal role to be played by our mental states. The conclusion is not that mental states do not exist, but rather that we have no reason to think that they have a causal effect.

The same argument applies more specifically to *our* reasons. For the purpose of this discussion, I understand reasons as facts which explain or favour other facts (or which explain both normative and non-normative facts). And I understand *our reasons* as pieces of knowledge or justified beliefs in these reasons. When *S* acts for reason *R*, *S* acts because *S* knows *R* or because *S* is justified in believing *R*. I make these plausible[18] assumptions for simplicity's sake. On this picture, if mental states have no causal role, nobody acts for a reason. It vindicates free will scepticism.

The worry is that if our reasons have no causal role on our conduct, the same applies for our beliefs, including our belief in free will scepticism. This Epicurean response was most famously made by Karl Popper (and has not been much discussed in the recent literature):

> [P]hysical determinism is a theory which, if it is true, is not arguable, since it must explain all our reactions, including what appears to us as beliefs based on arguments, as due to *purely* physical conditions […]. But this means that if we believe that we have accepted a theory like determinism because we were swayed by the logical force of certain arguments, then we are deceiving ourselves, according to

[18] See Hyman (2015) and Alvarez (2017).

> physical determinism […]. (Popper 1972, 223–224, emphasis is mine; see also Lockie 2018, chap. 10)

Thus, it is self-defeating to believe in free will scepticism because if it were true, it would follow that we do not believe it *for* a reason, since we believe *only* because of natural events. (Popper's use of 'purely' is supposed to mark an exclusion.)

Popper was right to maintain this connection between actions and beliefs. If we never act for reasons, we never believe for reasons. If my *giving* to charity being caused by a natural event is incompatible with (or if it excludes) my *giving* to charity for a moral reason, then my *believing* that I should give to charity being caused by a natural event is incompatible with (or it excludes) my *believing* it for a moral reason. For, in general, what counts as a reason to act can also count as a reason to believe. The fact that wealth is unequally distributed in our society is *both* a reason to *give* to charity and to *believe* that we should do so. And once we know that wealth is unequally distributed, then it becomes our reason to act and believe. What counts as a cause for a belief can also count as a cause for an action. The Wall Street Crash of 1929 caused many people to believe that they had lost their fortune; and it caused Roosevelt to propose his New Deal. And what counts as acting or believing *for* a reason must include some sort of causal connection between an agent's holding to that reason and her action or belief.

This clear but admittedly controversial picture leaves no room for an asymmetry that the free will sceptic could exploit in responding to the Epicurean. This suggests that the Epicurean has a sound argument:

> **The Reason Argument**
>
> If everyone acts because of natural events rather than for reasons, everyone believes because of natural events rather than for reasons.
> If everyone believes because of natural events rather than for reasons, no one believes rationally.
> *Therefore*, if everyone acts because of natural events rather than for reasons, no one believes rationally that free will scepticism is true.

I have just discussed the first premise, which is the typically Epicurean premise that connects conduct and beliefs. And the second rests on a natural connection between rationality and reasons and on the fact that by 'rather', I signal an incompatibility. As I indicated, it follows that believing

in free will scepticism on a certain ground is self-defeating, the ground being that we act because of natural events rather than for reasons. This is all I need to show for my purpose:

> **First lesson:** The free will sceptic cannot and should not rely on the doubtful claim that we act because of natural events rather than for reasons.

So free will sceptics should try, as we all should, to find a way out of the exclusion concern.

Since the problem is general and applies to all mental states, the free will sceptic will not be able to solve it by drawing distinctions between different kinds of explanations or reasons. True, she can insist—as did Ayer (1963, 266–267) and Wolf (1993, 72)—that something can be both caused and justified. If I ask you 'why do you believe that Italian is the most beautiful language?', you may *both* respond that it is because you spent some time in Italy (which merely explains your belief) or that it is more colourful than any other language (which justifies your belief). But this does not help to deal with the exclusion concern. This is because action or belief *for* a reason (*because* I am justified) must make room for the causal role of reasons; and this causal role is precisely what the exclusion concern targets.

4.2. Ownership

We now move to the second sceptical concern and its corresponding Epicurean argument. The concern is that even if we act for reasons (*contra* the exclusion concern), these reasons—and so our actions—are not truly *ours* since they can all be ultimately explained by facts or events *foreign* to us. Alternatively, even if some reasons (or desires and values) were ours, this would not mean much since our having these reasons would not ultimately be up to us (see Strawson 1994; Pereboom 2014a).

The Epicurean can respond to this challenge, again by connecting actions and beliefs. If our actions and our reasons for action are foreign to us, so are our beliefs. But then our beliefs do not stem from *our* reasons and threaten to be irrational.[19] The argument can be put in terms of sourcehood:

[19] See Kant (Groundworks 4: 448) for the claim that foreign influence is incompatible with the work of reason.

The Sourcehood Argument

If no one is the adequate source of their conduct, no one is the adequate source of their beliefs.
If no one is the adequate source of their beliefs, no one believes rationally.
Therefore, if no one is the adequate source of their conduct, no one believes rationally that free will scepticism is true.

The first premise is the typically Epicurean connection between conduct and beliefs. I shall argue that although the Epicurean is right to connect beliefs and actions in this way, she is wrong to think that moral free will (or responsibility) and rationality each requires sourcehood in the same sense. In a word, the *Sourcehood Argument* equivocates.

To be the adequate source of her conduct in the sense relevant to responsibility and free will, a subject must first be the *source* of her conduct. Her conduct must *stem from her*. Thus, it cannot be the direct result of an external factor such as someone pushing her. Nor can it be the direct result of an internal factor which is not the agent's, such as an uncontrollable impulse or a disease. For a similar reason, someone who acts under duress is not the adequate source of her action: this action does not really *stem from her* in the relevant sense. It somehow is the action *of the coercer* (Nozick 1997, 38).

But this is not enough. A child soldier who participates willingly in an act of war is still not the *adequate* source of his conduct. He did not make himself; and the commitments that he manifests in his violent behaviour are not fundamentally *his*. Some sceptics suggest that this is because the child is not *the source of the source* of his conduct (see Strawson 1994; Pereboom 2014a). To use Hurley's phrase (2003), the free will sceptic (and some incompatibilists) requires that to be responsible for Φ, someone be responsible for the causes of Φ. Now, to be responsible for Φ or to be the source of Φ, a subject does not need to be responsible for or be the source of everything that leads to Φ. The subject must be responsible for or be the source of what *determines* that she will Φ (see Istvan 2011). This still leads to a regress, unless the causal chain can stop in the agent's free act of will—which the sceptic denies. The subject's conduct is ultimately determined by something that she is not the source of—and so for which she is not responsible. The subject will never be the *ultimate source* of her conduct—she will never be truly responsible.[20]

[20] For a similar reasoning, see Galen Strawson (1994) and Istvan (2011).

Free will sceptics can say the very same thing about beliefs: to be the adequate source of them, we need to be the adequate source of their cause. But it leads again to a regression: for the causal chain continues up until a point where we no longer are the adequate source of the thing we are considering. Since this reasoning is just as plausible—no more, no less—for beliefs and for conduct, the first premise looks true: if no one is the adequate source of their conduct, no one is the adequate source of their beliefs.

Now, the second premise also seems true. Believing for good reasons or on adequate evidence is generally insufficient to believe rationally. For this, a subject must usually believe for good reasons *that are her own*—on *her* evidence. If Detective Chief Inspector Japp does not believe that there is an earring in the room, then it would be irrational at best of him to believe that the killer is a woman on the grounds that there is an earring in the room. Japp clearly needs to *acquire* the evidence or the reason—that he makes it *his*—and this in turn implies that he forms the relevant belief or that he comes to know the relevant fact.

For this, one might think, Japp needs again to believe on *his* evidence or for *his reasons*. And this might look like it causes a regress which shows that rationality, just like responsibility, is impossible. Yet this is mistaken. The fact that, to believe rationally, Japp needs to believe for his reasons does not imply that the reasons for which he believes must have been acquired because of a further reason that was his own. Why? There are two classic answers to this. The foundationalist (most notably Aristotle) claims that there might be self-evident reasons or special pieces of evidence that Japp acquires simply by paying attention to them. The existence of these things clearly has nothing to do with free will or determinism. The coherentist (such as Blanshard) claims that a series of belief can justify each other in a circle. Again, whether someone holds coherent beliefs has nothing to do with free will and determinism. The foundationalist can claim that while a responsible action requires a further responsible action *ad infinitum*, it is clear that a justified belief does not require a further justified belief *ad infinitum*. And a coherentist can claim that while justification can stem from the coherence of a set of beliefs, responsibility cannot stem from the coherence of a set of actions.

This means that, whether coherentist or foundationalist in nature, the structure of justification differs from the structure of responsibility. What is required for justification differs fundamentally from what is required for free will. Only the former can plausibly give rise to an infinite regression. We saw that there is a sense in which the concept of 'adequate sourcehood' is fitting in both the case of responsibility and of justification since each

implies a kind of ownership. But since the structure of justification is so different from the structure of responsibility, this strongly suggests that 'sourcehood' does not refer to the same property in each case or that the standards of adequacy differ fundamentally from premise 1 to premise 2.

The sceptic can therefore insist that the *Sourcehood Argument* equivocates: rationality requires sourcehood or adequacy in a different sense than moral responsibility (according to the sceptic). It is not, then, that 'adequate sourcehood' always means something different when it applies to actions and when it applies to beliefs. The Epicurean is right to maintain a symmetry between these things. Rather, 'adequate sourcehood' is different whether we are talking about what is necessary *for moral responsibility* and whether we are talking about what is necessary *for rationality*.

Despite having refuted the *Sourcehood Argument*, the free will sceptic has a concession to make. She must concede that when some compatibilist argues that we have free will because, roughly, we act for our own reasons, he is partly right. We *do* act for our own reasons, in the same sense that we believe for our own reasons. But this sourcehood, the sceptic must then argue, is different from the sourcehood necessary for moral responsibility. In a word:

> **Second lesson:** The free will sceptic must recognise the existence of some property of sourcehood, one that is necessary for people to believe and act for their own reasons and rationally.

The *Sourcehood Argument* fails but it urges the sceptic to explain why responsibility and justification use different notions or standards of sourcehood.

5. Leeway

As I indicated previously, some philosophers doubt the existence of free will on the grounds that we lack leeway. It suffices for our purpose to roughly indicate why. Typically, the leeway sceptic combines the Consequence Argument (see, e.g., van Inwagen 2017) and the Mind Argument (see, e.g., Pereboom 2014a). The Consequence Argument claims roughly that if determinism is true, then our actions are the consequence of the remote past and of the laws of nature. Thus, to act differently, we would need to change the laws of nature or the distant past, which is impossible. The Mind Argument says, crudely, that indeterminism is determinism plus chance, which is irrelevant to our abilities. Therefore, we lack the ability to do otherwise.

This suggests an Epicurean argument:

> If pessimistic determinism [free will scepticism] were true, no one could responsibly think that he had made a *wise* [or rational] *decision in believing* it. He had *no choice* but to believe it. (Dworkin 2011, 225, my emphasis)

The notion of choice is obviously tied to that of alternatives: Dworkin is not merely talking about the mental event of choosing, but the selection of one amongst several courses of action. Dworkin tells us, without much argument, that wisdom requires leeway.

A stronger philosophical treatment is given by Robert Lockie:

> [I]f determinism is true, then no-one can do otherwise and therefore *no-one may reason otherwise*. Assuming that the ability to reason otherwise is necessary for someone to be held epistemically irresponsible, no-one may then be held responsible for their intellectually wrong actions or unjustified, *irrational cognition*. But if no-one is responsible for their unjustified cognition, then no-one is epistemically justified either—in the intended, internalist sense. […] So one cannot be epistemically justified in claiming that determinism is true. (Lockie 2018, 183; my emphasis, premise numbering omitted)[21]

I shall argue that there is a simplified and less ambitious version of Lockie's argument. This argument does not suppose, controversially, that determinism is incompatible with leeway; it does not dip into the question of epistemic responsibilities; and it does not tackle the difficult topic of 'holding' responsible. My argument simply captures Lockie's insight that rationality ('internalist justification') requires leeway because leeway is required for obligations (responsibilities).

That leeway implies rationality cannot be established directly: it is false that if one believes a proposition rationally, one could have believed otherwise. A subject could believe a proposition rationally in the absence of any kind of leeway. Thankfully, most of us would be incapable of

[21] In this chapter, Lockie discusses two related arguments, but they are of a different kind. The first (179) concludes that determinists cannot accept deontic ethics; the second (about the lazy argument, 178) fundamentally makes the claim that in deliberating we must presume that we have options.

disbelieving that 2+2=4 and yet nothing impedes (directly) the rationality of our belief that 2+2=4. We need an indirect argument.

5.1. The Leeway Argument

The Epicurean can pursue an indirect strategy and claim that if one *irrationally* believes a proposition, one could have believed otherwise. This principle is more plausible. For, suppose that a subject believes that 2+2=5 and could not believe otherwise. Then, the Epicurean may still declare that this subject's belief is *not* rational: it is perhaps *arational*, but not *irrational*.

Why would the Epicurean say this? Irrationality, she can say, implies a failure to satisfy obligations to believe and if leeway does not exist, then these obligations do not exist either. Why does the inexistence of leeway imply the inexistence of obligations? The Epicurean says: first, if we can never do or believe otherwise, we never have the obligation to do or believe otherwise because 'ought' implies 'can' (henceforth OIC). Second, if we never have the obligation to do or believe otherwise, we have no obligations at all: we want to leave aside the abstruse view that we only have the obligation to perform the acts that we in fact perform and to believe the propositions that we in fact believe (this only make sense for a saint).

Once we have admitted that the inexistence of leeway implies the inexistence of irrationality, we should admit that it implies the inexistence of *rationality* in general. Rationality and irrationality go hand in hand. This explains why it is quite natural to think that animals and toddlers are *a*rational rather than *ir*rational. This does not have to mean that for one to rationally believe that 2+2=4 one is able to believe otherwise *in this situation*; it simply means that the existence of rationality in general implies the existence of irrationality in general.[22] We can make sense of the philosophical view that rationality *and* irrationality do not exist: rationality would be a property without extension (perhaps because there are no norms of rationality) and irrationality would not exist because irrationality is *absence of rationality where rationality could apply*. Everything would be *a*rational because arationality means the absence of rationality where rationality could not apply (e.g. the table is arational). But to make sense of the view that rationality exists, but not irrationality, we would need to imagine that if rationality can apply, then it exists. This only works in an ideal world, a place we do not live in.

[22] This argument might recall the theistic argument that the existence of evil implies the existence of the good.

It is now helpful to summarise the argument:

> **The Leeway Argument**
> If no one has leeway, there are no obligations.
> If there are no obligations, no one believes irrationally.
> If no one believes irrationally, no one believes rationally.
> *Therefore*, if no one has leeway, no one believes rationally that free will scepticism is true.

Here again we have a typically Epicurean connection between conduct and beliefs in the two first premises (since 'no one has leeway' is at least partly about conduct).

If the Leeway Argument is sound, it follows that it is self-defeating to believe that free will scepticism is true on the grounds that we have no leeway. The presumed truth of free will scepticism ensures that either it is believed on false grounds (since we *have* leeway) or that it is not believed rationally (since we have no leeway and thus cannot believe rationally).

It is worth insisting at this point that, although the sceptic could admit the existence of leeway and find other grounds for her doctrine, this is a major consequence. As we just saw, some of the main arguments for free will scepticism feature leeway. And, although many sceptics[23] and non-sceptics[24] are now sourcehood theorists, they often concede in passing that, in a sense, we lack leeway. If the Leeway Argument is sound, this is problematic; belief leeway needs to be maintained for rationality.

This means that the sceptic should try to find a way to refute the Leeway Argument. I suggest that she abandons OIC, although this might be mistaken. Let me simply point out that OIC is popular amongst free will sceptics (see, e.g., Pereboom 2014b, 222; cf. Waller 2011). Some sceptics have appealed to it in defence of another principle that is sometimes used by free will sceptics, the principle that we are only responsible if we could have done otherwise.[25] And if free will sceptics were to concede that we *can* do otherwise, it might be harder to argue that we are not ultimately the source of our actions.

Instead of making this costly concession, the sceptic might want to follow a third way and declare that the Leeway Argument equivocates on the sense of 'obligation'. There are two ways to argue for this. The first is to

[23] See Pereboom (1995, 27; 2014a, 138; 2014b, 221).
[24] This might include semi-compatibilists like Fischer (1994).
[25] On this, see e.g. Widerker (1991) and Copp (2008).

oppose obligations to believe and obligation to act and insist that the second but not the first requires leeway. The second is to oppose epistemic obligations and moral obligations and maintain that the second but not the first requires leeway. I consider each in the following subsections.

5.2. Obligations to Believe, Obligations to Act

We have some power over our conduct and so while it is controversial that an obligation to act implies 'can', it remains plausible. By contrast, obligations to believe, one might suggest, cannot imply 'can' because we have obligations to believe, but we do not control directly our beliefs.[26] So, although we might have thought with the Epicurean that all obligations were similar in their requirements, this is incorrect: obligations to believe and to act are made of a different fabric.

While tempting, this counterargument is unavailable to our sceptic (and so there is no need to consider whether we directly control our beliefs). This is simply because the sceptic who accepts that we lack leeway, following the same reasoning, would have to conclude that the moral 'ought' does not imply 'can' either, since she denies the existence of leeway in conduct. But then the strategy is no longer a third-way escape focussing on the asymmetry between two kinds of obligations to preserve OIC. It throws OIC under the bus.

5.3. Epistemic Obligations, Moral Obligations

The contrast between epistemic and moral obligations might look more promising. As Clifford insisted, there are moral obligations to act, but also to believe. Similarly, there are epistemic obligations to believe, but there might also be epistemic obligations to act. So the epistemic/moral distinction cuts across the to believe/to act distinction. If I must choose between happily believing against my evidence and depressingly believing on my evidence, it is possible that the two kinds of normativity pull me in different direction. We can see this if we consider a case where a doctor could gather more evidence for the efficacy of a new vaccine or spend this time to administer it to patients. The general idea is that there are two kinds of normativity which aim at two kinds of value and which are subject to different requirements. Perhaps the anti-Epicurean can use this picture.

Her strategy would be to argue that moral, but not epistemic, obligations require alternatives. For this, the anti-Epicurean will try to show that epistemic normativity is like other 'minor' kinds of normativity. In fact, in

[26] This is precisely what Alston (1989) advocates.

general, certain kinds of obligations, like obligations of etiquette and professional obligations, do not imply 'can.' The Dean of the Faculty might have an obligation to attend a meeting even if he cannot be present—he might have been in a car accident. He thus fails to satisfy an obligation, although he is excused. To be sure, his work-description does not specify that his obligations are suspended if he is unable to meet them, at least, not in all cases where he lacks the relevant abilities. Similarly, etiquette certainly recommends that a guest show appreciation for a meal, but this might be impossible if a food allergy forces her to rush to the bathroom. An obligation was infringed, despite the excuse. And the epistemic 'ought' might be very similar to etiquette and professional obligations as regards 'can': "our friend in his tinfoil hat can't make himself stop overtly believing contradictions" (Carr 2015, 752). But, surely, our friend in his tinfoil hat is not believing rationally and therefore he infringes an obligation of rationality.

But then, can't we say the same thing of moral obligations? There is an analogous character to our friend in his tinfoil hat—the inexorably evil man with his dictator cap. Surely, he infringes his moral obligations—and so the moral 'ought' does not imply the moral 'can'. Remember however that the anti-Epicurean strategy here is to keep '"moral ought" implies "can"' and reject '"epistemic ought" implies can'. So she must insist that somehow the dictator does not have the *true* moral obligation to refrain from 'dictating.'

She can make the following case. True obligations imply can (perhaps because their infringement is blameworthy), and other obligations—such as professional obligations, chess obligations, and epistemic obligations—are just rules that it would be ridiculous to insist 'imply can'. What distinguishes our friend in his tinfoil hat from our foe in his dictator cap is that the first fails to follow epistemic *rules* whereas the second is not under moral *obligations*.[27] Satisfying epistemic rules or rules of etiquette then has no intrinsic significance: it does not matter, in and of itself, whether one has rational beliefs or not. The significance of following those rules is quite unlike, for instance, the significance of finding meaning in one's life or of respecting others. Hence, it is no surprise if general principles like OIC do not apply to epistemic obligations (or rules).

I am disinclined to admit this response because I think that at least some epistemic statuses, as opposed to statuses of etiquette, bear intrinsic or final significance or value. To be fair, some false beliefs seem too negligible to matter for their own: whether I truly believe that there is an odd number of

[27] See Côté-Bouchard (2017).

grains of sand on this beach is negligible. And yet, it seems that *in general* having true beliefs—but not following rules of etiquette—*is* intrinsically or finally significant or valuable. We think it is intrinsically or finally valuable to be connected with the world—we want to be guided by the truth. And this connection, just is an epistemic connection (see Hyman 2015, 209).

Now, to be sure, the proposition that *knowledge* or *rationality* is finally or intrinsically valuable does not entail that *following epistemic* obligations is. But this seems enough to close the gap between ethics and epistemology and to widen the gap between epistemology and etiquette. (After all, perhaps *following moral rules* has no intrinsic value. Perhaps it is only resulting states of affairs that have intrinsic value.)

We should add that the anti-Epicurean is in no position to insist too much on individual cases of valueless knowledge or of true belief. For a similar problem will appear in the case of morality. Imagine that we have a pair of cases of identical situations of horrible actions except that in the first, but not in the second, the villain could not have done otherwise, and so infringes no moral obligation. Just like it seems indifferent whether we have a true belief or an item of knowledge about the number of grains of sands, it seems indifferent whether we are in the first or in the second case—whether the villain is infringing an obligation. We should not infer from these sorts of cases that there is no real epistemic or moral normativity.

It seems, then, that sceptics have a solid case for the symmetry of moral and epistemic obligations regarding leeway. This takes us back to the two other options, which are costly to the sceptic. Thus, we can say:

> **Third Lesson:** The free will sceptic must either: (1) accept that we may have leeway; or (2) reject OIC.

None of these avenues is blocked to the sceptic, but, as we have seen, each is at least *prima facie* very costly.

6. Conclusion

Epicurus' argument was *prima facie* tempting. There seemed to be something suspicious about claiming, on the one hand, that we have no free will but, on the other hand, engaging into fierce deliberation about what would be best and rational for us to achieve and believe.

Trying to give the best reading of Epicurus' argument helped us to identify some of the arguments underlying this initial sentiment. There was *something* to it: not just a confused feeling, but sourcehood-related and leeway-related arguments concluding that to believe in free will scepticism is self-defeating. Yet none of the considered arguments was fully successful. That's a major point in favour of the free will sceptic.

Still, our discussion left the sceptic with three lessons:

> **First lesson:** The free will sceptic cannot and should not rely on the doubtful claim that we act because of natural events rather than for reasons.
> **Second lesson:** The free will sceptic must recognise the existence of some property of sourcehood, one that is necessary for people to believe and act for their own reasons and rationally.
> **Third Lesson:** The free will sceptic must either: (1) accept that we may have leeway; or (2) reject OIC.

These lessons should be added to the sceptic's breviary. What is their significance? The third lesson makes it much harder to be a free will sceptic of the leeway kind; and the first lesson deprives the sceptic of an important argument (from the exclusion concern). The obvious way forward for free will scepticism seems therefore to revolve around the ownership concern, according to which we cannot be responsible for our conduct because even if we act for our reasons, desires, or values, they are not as much ours (or significantly ours) as we thought; in fact, we own them because of factors lying outside of our control. This might give confidence to the free will sceptics, who have mostly turned to this sort of view in the past decades (Strawson 1994; Pereboom 2001; Waller 2011; Pereboom 2014a). Yet sourcehood sceptics were mistaken in thinking that since leeway was not the crux of responsibility, they could deny the existence of leeway on top of sourcehood at no additional cost.[28]

Although, I did not have the chance to tackle this question seriously (see §2.2), I believe it would be a mistake for the sceptic to accept that her doctrine is self-defeating and insist that at least she has a true belief. After all, what made her view attractive was that she gave us powerful *reasons* to believe that some of our institutions are *unjustified* and some of our attitudes are *irrational*. To admit that being a sceptic is itself neither rational nor reasonable would undermine this. Free will scepticism is a philosophical doctrine supported by arguments, not a creed supported by hopes.

[28] See Pereboom (1995, 27; 2014a, 138; 2014b, 221).

Acknowledgements

I wish to thank Ralf Bader, Charles Côté-Bouchard, Tanya Goodchild, Alison Hills, John Hyman, Clayton Littlejohn, and two anonymous reviewers of this journal for their invaluable comments on previous versions of this essay.

REFERENCES

Alston, W. 1989. *Epistemic Justification: Essays in the Theory of Knowledge*. Ithaca; London: Cornell University Press.

Alvarez, M. 2017. Reasons for action: Justification, motivation, explanation. *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta. Accessed September 22, 2019. https://plato.stanford.edu/entries/reasons-just-vs-expl/

Annas, J., W. Taylor, and D. N. Sedley. 1983. *Oxford Studies in Ancient Philosophy.* Oxford: Clarendon Press.

Ayer, A.J. 1963. The Concept of a Person: And Other Essays. London: Macmillan.

Bailey, C. 1926. *Epicurus: The Extant Remains*. Oxford: Clarendon Press.

Bobzien, S. 2000. Did Epicurus discover the free-will problem? *Oxford Studies in Ancient Philosophy* 19: 287–337.

Carr, J. 2015. Don't stop believing. *Canadian Journal of Philosophy* 45: 744–766.

Castagnoli, L. 2007. Everything Is True, Everything Is False: Self-Refutation Arguments from Democritus to Augustine. *Antiquorum Philosophia* 1: 11–74.

Chevarie-Cossette, S.-P. 2019. Self-defeating beliefs and misleading reasons. *International Journal of Philosophical Studies* 27: 57–72.

Chevarie-Cossette, S.-P. Forthcoming. Knowing about responsibility: A trilemma. *American Philosophical Quarterly*.

Copp, D. 2008. "Ought" implies "can" and the derivation of the Principle of alternate possibilities. *Analysis* 68: 67–75.

Côté-Bouchard, C. 2017. Belief's Own Metaethics? A Case against Epistemic Normativity. *KCL doctoral thesis*.

Cuneo, T. 2007. *The Normative Web: An Argument for Moral Realism*. Oxford: Oxford University Press.

Dworkin, R. 2011. *Justice for Hedgehogs*. Harvard University Press.

Fischer, J.M. 1994. The Metaphysics of Free Will: An Essay on Control. Oxford: Blackwell.

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

Frede, M. 2011. *A Free Will: Origins of the Notion in Ancient Thought*. Berkeley; London: University of California Press.
Hinman, L. 1979. How not to naturalize ethics: The untenability of a Skinnerian naturalistic ethic. *Ethics* 89: 292–97.
Hurley, S. 2003. *Justice, Luck, and Knowledge*. Cambridge: Harvard University Press.
Hyman, J. 2015. *Action, Knowledge, and Will*. Oxford: Oxford University Press.
Inwagen, P. 2017. *Thinking about Free Will*. Cambridge: Cambridge University Press.
Istvan, M. 2011. Concerning the resilience of Galen Strawson's Basic argument. *Philosophical Studies* 155: 399–420.
James, W. 1912. The Will to Believe: And Other Essays in Popular Philosophy. Auckland: Floating Press.
Kim, J. 2007. *Physicalism, or Something near Enough*. Princeton: Princeton University Press.
Lockie, R. 2018. Free Will and Epistemology: A Defence of the Transcendental Argument for Freedom. New York: Bloomsbury Academic.
Mackie, J.L. 1964. Self-refutation: A formal analysis. *The Philosophical Quarterly* 14: 193–203.
Nozick, R. 1981. *Philosophical Explanations*. Oxford: Clarendon Press.
Nozick, R. 1997. *Socratic Puzzles*. Cambridge: Harvard University Press.
Olson, J. 2014. Moral Error Theory: History, Critique, Defence. Oxford University Press.
Page, C. 1992. On being false by self-refutation. *Metaphilosophy*. 23(4): 410–26.
Pereboom, D. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.
Pereboom, D. 2014a. *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.
Pereboom, D. 2014b. Responses to John Martin Fischer and Dana Nelkin. *Science, Religion & Culture* 1(3): 218–25.
Popper, K. 1972. Objective Knowledge: An Evolutionary Approach. Oxford: Clarendon Press.
Sartorio, C. 2016. *Causation and Free Will*. Oxford: Oxford University Press.
Shoemaker, D. 2015. *Responsibility from the Margins*. Oxford: Oxford University Press.
Slagle, J. 2016. The Epistemological Skyhook: Determinism, Naturalism, and Self-Defeat. London and New York: Routledge.
Strawson, G. 1994. The impossibility of moral responsibility. *Philosophical Studies* 75(1/2): 5–24.

Vihvelin, K. 2013. Causes, Laws, and Free Will: Why Determinism Doesn't Matter. Oxford: Oxford University Press.
Waller, B. 2011. *Against Moral Responsibility*. The MIT Press.
Widerker, D. 1991. Frankfurt on "ought implies can" and alternative Possibilities. *Analysis* 51: 222.
Wolf, S. 1993. *Freedom within Reason*. Oxford: Oxford University Press.

EuJAP | Vol. 15 | No. 2 | 2019
UDK: 123.1 165
https://doi.org/10.31820/ejap.15.2.4

HOW DO WE KNOW THAT WE ARE FREE?

Timothy O'Connor
Indiana University

Original scientific 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 depends causally on my having had prior experiences and conceptual

[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).

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

Schurger, A., J. D. Sitt, and S. Dehaene. 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.

EuJAP | Vol. 15 | No. 2 | 2019
UDK: 123.1:165.023.2
https://doi.org/10.31820/ejap.15.2.5

THE CONCEPTUAL IMPOSSIBILITY OF FREE WILL ERROR THEORY

Andrew J. Latham
The University of Sydney

Original scientific article – Received: 30/04/2019 Accepted: 13/11/2019

ABSTRACT

This paper argues for a view of free will that I will call the conceptual impossibility of the truth of free will error theory - the conceptual impossibility thesis. I will argue that given the concept of free will we in fact deploy, it is impossible for our free will judgements—judgements regarding whether some action is free or not—to be systematically false. Since we do judge many of our actions to be free, it follows from the conceptual impossibility thesis that many of our actions are in fact free. Hence it follows that free will error theory—the view that no judgement of the form 'action A was performed freely'—is false. I will show taking seriously the conceptual impossibility thesis helps makes good sense of some seemingly inconsistent results in recent experimental philosophy work on determinism and our concept of free will. Further, I will present some reasons why we should expect to find similar results for every other factor we might have thought was important for free will.

Keywords*: Free will, error theory, conceptual impossibility, conditional concept, experimental philosophy*

1. Introduction

Strictly speaking, transcendental arguments are arguments that attempt to show that *X* is a necessary precondition for the possibility of *Y* and hence since actually *Y*, therefore actually *X*. Immanuel Kant (1781/1787) is, of

course, the most famous defender of arguments of this kind. We can find examples of this kind of argument throughout many different domains of philosophy. One recent example involves an objection to certain approaches to quantum gravity in the philosophy of time. These approaches are said to be timeless, since they deny there exists any ordered series of events that are temporally or causally connected to one another. However, a necessary precondition to even entertain these theories is having contentful mental states. But having contentful mental states requires causal connections between at least some of our mental states and states in the world those states are about. So, we are only able to entertain these theories if in fact they are false (Braddon-Mitchell and Miller 2018). Of course, one response to a transcendental argument is to just deny what the proponent takes to be undeniable. For instance, in philosophy of mind, a proponent of eliminative materialism, of the kind defended by the Churchlands (1981; 1986), can just deny that you need to have beliefs (rather than other neuroscientific states) in order to argue that there are no beliefs.[1]

A transcendental argument for free will would proceed by showing that the necessary precondition for the possibility of some way things actually are—for instance, our being agents, or deliberators, or the kinds of things that can ask questions about free will—is there being free will. It then follows that since we are such things, there is free will. Robert Lockie (2018) does just this in his new book *Free Will and Epistemology*: If our having libertarian free will (free will incompatible with determinism) is a necessary precondition for the possibility of our having any justified beliefs, then if we believe that we do not have free will, either this belief must be unjustified, if it's true, or if justified, it must be false. In this paper, I will run a different line of argument to the conclusion that we have free will. Roughly, for now, the idea will be that most of our actions being free is a necessary precondition for understanding our ordinary practices as being non-defective, and as they are not defective, we have free will.

In this paper, I will argue that our concept of free will cannot do the job it is supposed to do, and that concept fail to be satisfied. That's because most of our actions being free is a necessary precondition for understanding our ordinary free will practices as being non-defective. These practices involve drawing certain kinds of distinctions between different kinds of actions that we track with our talk of free and unfree. We distinguish actions performed while being coerced, from those performed while fulfilling our desires, and actions performed in the grips of a mental illness, from those performed after some long effortful deliberation. It's important to note that what I am

[1] Thanks to Kristie Miller for bringing these cases to my attention.

intending to pick out in discussing our free will practices is much wider than our moral responsibility practices. Consider for a moment certain kinds of advertising which push us towards choosing one option over another. While these advertisements might impact our behavior in a predictable manner, they do so in a way which is not *mentally mediated*. That is, the advertising seems to impact behavior via sub-personal level processes which are not consciously available to the deliberator. What is important is that the reason we don't like these kinds of advertising pushes is *not* because we think they undermine our moral responsibility, but because they seem to impact our free will in a manner we don't like. For the purposes of this paper I am going to assume we could not engage in these practices without making these kinds of distinctions, and further, that these practices cannot and should not be revised. The argument for this claim about our practices is a job for another paper. Given that these practices are not defective, then, I argue, we have free will. It is, as it were, *conceptually impossible* for us to deploy the concept of free will that we do, and the world fail to satisfy that concept.

An analogy: one might argue that our concept of ordinary objects such as trees, rocks, and so on, are such that even if it turned out that we are living in a computer simulation, or some demon's brain, it will still turn out that there are trees and rocks. We might discover that their underlying nature is surprising, but not that they don't exist (Chalmers 2005). If our concept of tree was something like: whatever thing it is with which I am causally connected, when I have mental states of *this* kind, then, it would simply turn out that if our world is a computer simulation, trees *are* parts of such simulations. What trees are fundamentally made of turns out to be different than we originally supposed, but that doesn't mean there are no trees.

I will argue that it cannot be that we deploy the concept of free will that we do, and it turn out that actually we are systematically mistaken about which actions are free, and which actions are unfree. Of course, the idea that there could be such concepts might seem puzzling, so in §2 I will outline and defend the conceptual impossibility thesis. Then in §3 I will show how taking seriously the conceptual impossibility thesis reconciles some apparent inconsistencies in the extant empirical evidence regarding our concept of free will, and determinism. In §4 I will give reasons why we should think that the finding that determinism doesn't matter for our having free will, should generalize to other factors people have thought were important for free will. Finally, in §5 I will conclude.

2. The Conceptual Impossibility Thesis

Before I outline the conceptual impossibility thesis in more detail, some clarifications are in order. The conceptual impossibility thesis is the thesis that given the content of the concept of free will that we, the folk, in fact deploy, it cannot be that the concept is *systematically* misapplied. That is, it cannot be that we are systematically mistaken about which actions are free and which are unfree. Two things are noteworthy here. First, the concept with which I am interested is the *folk* concept of free will. There might be *philosophical* re-conceptions of free will which have quite different content from the folk concept, and I will make no attempt to consider such concepts here. Second, the conceptual impossibility thesis is a thesis about *systematic* error. It is not the thesis that none of our judgements about which actions are free (or not) are false. It is consistent with the conceptual impossibility thesis that some of our judgements about which actions are free (or not) are mistaken.

Why would one accept the conceptual impossibility thesis? Let's call a judgement of the form 'action A is free' a *positive* judgement, and a judgement of the form 'action A is unfree' a *negative* judgement. I will argue that the content of our folk concept is *something* like the following: free will is whatever thing there is in the world which most of our positive judgements track. In this regard, I argue that our concept of free will has a content, which is such that however our world turns out to be, most of our free will judgements (both positive and negative) will be vindicated. The only way this could fail to be is if there were *nothing at all* in common between most of the times we judge that an action is free, and most of the times we judge that an action is unfree, such that we are not tracking anything at all, for there is nothing there to be tracked. But this is clearly not the case: there *are* such similarities. Even free will error theorists don't think that there are no such similarities; they simply think that those similarities are not, in fact, sufficient to vindicate our positive free will judgements.

But why think that the content of our concept is as I suggest?

Consider the kinds of cases that we ordinarily judge positively to be free, and judge negatively to be unfree. Ordinarily, we make positive judgments regarding cases where we are act in accordance with our reasons, in fulfilling our desires, after having mentally simulated numerous courses of actions and their projected outcomes, and so on. Conversely, we make negative judgments regarding cases where we are bound-up, or being coerced and manipulated, or caught in the grips of a psychological or physiological illness, and so on. Of course, neither list is exhaustive of all

the kinds cases that we judge positively and negatively. For the moment, I simply want to roughly flag the kinds of cases that we ordinarily think of as free and unfree.

One way of characterizing the problem of free will is as the worry that there is no metaphysical difference between the cases where we judge actions to be free, and those we judge to be unfree. That would seem to be the case were we to discover that some fact that characterizes those actions we currently class as unfree, turns out to be true of *all* our actions (Dennett 1984; 2013). For instance, if it turned out that all our actions are coerced, or manipulated, or in the grips of psychological or physiological illness, then *prima facie* this would seem to be the discovery that none of our actions are free.

Let us focus, for the moment, on one important metaphysical factor relevant for free will: determinism. Philosophers have traditionally thought that consideration of determinism is important for free will, and so it has received the most empirical attention in experimental philosophy. In §3 I will turn to the empirical data on the relationship between the folk concept of free will and determinism. In §4 I will give some good reasons to think that the lessons of the conceptual impossibility thesis generalize to all other relevant metaphysical facts as well.

For now, suppose we only judge actions to be free if they are not determined. Then indeterminism is necessary for our concept of free will to be satisfied, (as is commonly supposed),[2] and if we discover that determinism is true, then we discover that there is no free will. Notice, though, that if there's no free will, then *all* our actions are akin to being bound-up, or coerced and manipulated, or caught in the grips of a psychological or physiological illness. That, however, seems wrong. Even if there is no deep *metaphysical* difference between the cases, we judge to be free, and those we judge to be unfree, we still want our actions to be like the ones that we ordinarily think of as free. After all, even if, with respect to some particular metaphysical matter of fact, there is no difference between these actions, there still seem to be other relevant differences that we want to track with our talk of free and unfree action. We want to normatively evaluate actions—whether this be moral or prudential evaluation—and to do that we want to distinguish actions that are performed while being coerced and manipulated, from those that are not,

[2] See e.g. Ekstrom (2002), Kane (2005), O'Connor (2000), Pereboom (2001), Pink (2004), Strawson (1986), van Inwagen (1993) to name a few. Contra this some theorists such as Eddy Nahmias (2011) think the folk concept of concept is a compatibilist one and that incompatibilist judgments arise out of people misunderstanding the implications of determinism for free will.

and actions performed while in the grips of a psychological or physiological illness, from those that are not, and so on. Regardless of whether determinism is true, we can be expected to care whether our friend stood on our foot because, having deliberated about it, she decided this is what she wanted to do, and proceeded to do it, or because she was pushed over by the person next to her. *Mutatis mutandis* for all these kinds of cases.

So there seems to be a concept of free will that tracks superficial differences between the cases we judge to be free, and the cases we judge to be unfree. For ease of explication I will call this a *social kind concept*.

One might, however, object. Consider for the moment a potentially analogous case involving water and ice. According to the story I have provided so far there are *two* different social concepts. The *social* concept of water, which is sensitive to the stuff that fills the oceans, flows through the rivers, falls from the sky whenever it rains, and so on, and the *social* concept of ice which is sensitive to the stuff found in glaciers, around the poles of the Earth (for now), falls from the sky as hail, and so on. Yet while perhaps once we thought that water and ice were different kinds of things, as a result of scientific investigations we have discovered that there is no deep metaphysical difference between water and ice: they are both H_2O. So, we now believe there is only one *natural kind concept*, which both water and ice fall under.

Surely, we should expect the same thing to occur in the case of free will: discovering some deep metaphysical fact that characterizes actions we currently judge to be unfree, to be shared with actions we judge to be free, gives us warrant to conclude that both sets of actions are of the same *metaphysical* kind, and that both are unfree. For example, if determinism is in fact true, and we judge that such a metaphysical fact makes actions unfree, then we should judge that none of our actions are free.[3]

Thus, there seems to be a concept of free will that tracks some deep metaphysical feature of our actions. I will refer to the concept of free will that is relevantly similar to a natural kind concept, a *metaphysical kind* concept.

The idea that free will might be a natural kind has been expressed in the free will literature before (Heller 1996; Deery 2019). Such a view is a natural extension of the paradigm-case view advanced by Antony Flew (1955), who suggested that the meaning of ‘free will’ is fixed by the

[3] Thanks to David Braddon-Mitchell for the Ice and Water case.

paradigm cases.[4] However, if it's a conceptual constraint that something falls under the concept of free will only if that thing forms a natural kind, then the meaning of 'free will' is fixed by whatever natural kind is uniformly in common between all (and only) the paradigm cases.[5]

One consequence of thinking of free will as a natural kind is that it admits a family of views which vary according to what you think is in common between all the paradigm cases. For instance, on the one hand, free will might form a metaphysical kind and so carve nature at its joints. This seems to be the case when we think that free will is whatever allows our actions to be *indeterministic*, whilst not being *merely chancy*. On the other hand, free will might form a psychological, functional or social kind. While these latter kinds do not carve nature at its joints, they nevertheless carve nature up in a useful fashion. Perhaps free will is a psychological capacity or suite of psychological capacities, or perhaps free will is just the practices themselves of judging certain actions to be free and unfree. Finally, and most permissively, free will might just be whatever is a member of the set of paradigm-cases. On this view free will could be anything at all.

It is my view that we should treat this family of natural kind views as a kind of prioritized hierarchy.[6] By that I mean that if the metaphysical kind is there and in common between the paradigm cases, then that's what free will is and necessarily so. Else, if the psychological kind is there and in common between the paradigm cases, then that's what free will is, and necessarily so, and so on. Then perhaps, finally, if there is no natural kind in common between the paradigm cases, then free will just is the paradigm cases. While I think that there is something in common between the paradigm cases I am not taking a stand in this paper on exactly what that is. Further, I am not advocating that it is possible for anything at all to count as free will which would seem to be the case if there is nothing at all in common between the paradigm cases, aside from being a member of the set of paradigm cases. While I think that it's open for someone to think that, it is not my view.

For the ease of ongoing discussion I will restrict myself to just the social and metaphysical kinds. Given these two apparent concepts of free will, there are two conceptual impossibility theses: one *weak* and one *strong*.

[4] Thanks to an anonymous referee for making me aware of this existing and growing literature.

[5] Though for arguments against the paradigm-case view and free will as a natural kind, see van Inwagen (1983) and Daw and Alter (2001).

[6] I will have much more to say about this kind of prioritized hierarchy when I come to discuss the idea of the folk concept of free will being a conditional concept with respect to determinism in §3.

The *weak conceptual impossibility thesis* is that the social concept of free will and the metaphysical concept of free will are both important. If some underlying metaphysical feature is missing (i.e. determinism is true) then on the metaphysical concept of free will, error theory will be true. However, according to the weak conceptual impossibility thesis the social concept of free will is also important, and on the social concept there will be free will regardless. The conceptual impossibility thesis is true of the social concept. The *strong conceptual impossibility thesis* is that while both the social concept and metaphysical concept exist, it's only the social concept that matters, so the conceptual impossibility thesis is true of the concept that matters. Let me elaborate on both these theses.

2.1. The Weak Conceptual Impossibility Thesis

There are two apparent concepts of free will: a metaphysical concept which is open to the possibility that there is no free will (analogous to the discovery that since ice is just H_2O, in some deep sense there is no ice) and a social concept according to which as long as there *are* differences between paradigm cases we judge to be free and paradigm cases we judge to be unfree, this guarantees there is free will. On the weak conceptual impossibility thesis, both concepts are needed, and the social concept is guaranteed to be satisfied.

But what do I mean when I say both concepts are needed? Well the fact that water and ice are both H_2O plays an important explanatory role in our best scientific theories; such as why ice and water exhibit the same chemical properties. So, there is an important sense in which there is not both water and ice, there is just H_2O. Perhaps philosophers, too, will conclude that there's no metaphysical difference between those cases that we ordinarily judge to be free, and those we judge to be unfree. However, aside from generating an apparent problem for free will, I am not sure what purpose we have for taxonomising our actions according to their deep metaphysical nature. For instance, what is gained by classifying our ordinary actions by the lights of determinism? I will return to this point shortly when I describe the strong impossibility thesis. I leave it open, here, that there could be good reasons for classifying our actions according to their metaphysical nature (i.e. determinism), and thus to in some sense collapse the distinction between free and unfree actions on the metaphysical concept.

Even if we do so, however, there is clearly some relevant distinction between the actions we judge to be free, and those we judge to be unfree. To see this, return to the case of water and ice. Suppose we agree that there is no metaphysical difference between water and ice, and hence that in

some good sense we can collapse the distinction between them. Nevertheless, there is a clear sense in which despite this, there *is* both water and ice despite there being no metaphysical difference between them. That's because we care about the role the superficial differences between water and ice plays in ordinary matters. If I am thirsty and ask for a glass of water at a restaurant, I would be amused to receive a glass filled with ice.

Similarly, even if there is no deep metaphysical difference between actions we judge to be free, and to be unfree, we still care deeply about whether actions fall into one, or instead the other, category. We care whether or not we act for our reasons, in order to fulfil our desires, or after some process of deliberation as opposed to being bound-up, coerced and manipulated, or caught in the grips of a psychological or physiological illness. What this social concept of free will tracks then, is *whatever it is* which vindicates this difference.

The weak conceptual impossibility thesis holds that the distinction between free and unfree actions is like the distinction between water and ice. Just as there are two ways of thinking about water and ice, there are two ways of thinking about free and unfree action. On the metaphysical concept, we group the cases according to their metaphysical nature, and so decide that there are no free actions if determinism is true. This is analogous to the sense in which there is not water and ice, there is only H_2O. On the social concept we group the cases according to some, perhaps more superficial, difference between them, a difference that we care about for our ordinary purposes. This is analogous to the sense in which we ordinarily treat water and ice as distinct despite there being no metaphysical difference between them. That's because what we are often just as, if not more, interested in, is the role such distinctions play in ordinary matters, and not their deep metaphysical nature. So, while error theory is true of our metaphysical concept of free will, the conceptual impossibility thesis is true of our social concept of free will.

2.2. The Strong Conceptual Impossibility Thesis

What of the strong conceptual impossibility thesis? According to that thesis, while there are two concepts of free will, only the social concept *matters* for any important purposes. In the water and ice case, the metaphysical concept on which despite superficial differences, both water and ice are H_2O, plays an important role in our best scientific explanations in the chemical sciences. That's why the metaphysical concept matters. But there seems to me to be nothing analogous in the case of free will that justifies taking seriously the idea that just because something about every

free action turns out to be like the unfree actions, that that feature is crucial for freedom. The strong conceptual impossibility thesis says that the free and unfree distinction is not like the distinction between water and ice because while we certainly care about the superficial differences between those actions we judge to be free and unfree, there's nothing analogous to the chemical sciences which justifies taxonomising our ordinary actions according to deep metaphysical similarities. Error theory might be true on the metaphysical concept of free will, but no one ever cared about that concept because it doesn't matter for any of the purposes for which we deploy that concept. So, on the only concept that matters, the social concept, the conceptual impossibility thesis is true.

In the next section I will show how the conceptual impossibility thesis has important consequences for the interpretation of extant empirical work on our folk concept of free will and its relationship to the thesis of determinism. Then later, I will give some reasons to think that all factors that we might have thought mattered for free will (such as determinism) *don't*.

3. Experimental Philosophy, Determinism and the Folk Concept of Free Will

One metaphysical factor that many people have supposed matters for free will is determinism. The thesis of determinism holds that the entirety of particular facts about the past, in conjunction with the laws of nature, entails every truth about the future. Is our concept of free will compatible with determinism being true? Compatibilists answer affirmatively. According to them, if determinism is true then provided agents have some preferred set of abilities, which vary according to the version of compatibilism at issue, then free actions are those produced by those abilities. For ease of explication I will refer to whatever the abilities are that when exercised in the production of an action makes that action free according to compatibilism: *compatibilist powers*. Conversely, incompatibilists take it to be a necessary condition for our having free will that indeterminism is true. *Libertarians* are incompatibilists who think there is free will. Call whatever the abilities are that when exercised in the production of an action makes that action free according to libertarianism: *libertarian powers*.

If the conceptual impossibility thesis is true, then the folk concept of free will must be compatible with determinism. But, while it's often been assumed that the folk concept of free will is an incompatibilist one, there is excellent evidence from experimental philosophy that the folk concept

is a compatibilist concept (e.g., Nahmias et al. 2005; 2006) and also that it is an incompatibilist concept (e.g., Nichols and Knobe 2007).

How should we make sense of this apparent inconsistency? Roskies and Nichols (2008; though see also Björnsson 2014; Latham 2019) noticed a difference in the experimental materials used. While Nahmias and colleagues situated some of their determinism vignettes in the *actual* world, Nichols and colleagues situated them in *hypothetical* worlds. In order to confirm their suspicion that participants' free will judgements to deterministic vignettes differed as a result of where they were being evaluated, participants were evenly split between considering deterministic vignettes in the actual world or in some other hypothetical world. Consistent with the authors' hypotheses, where the deterministic scenario was situated significantly impacted participants' free will judgements. Participants' free will judgements were significantly higher when the deterministic vignette being evaluated was in our own world relative to when the deterministic vignette being evaluated was in some hypothetical world.

3.1. Determinism and a Conditional Concept of Free Will

Roskies and Nichols (following Braddon-Mitchell 2003; though see also Latham 2019) argued that these results suggest that the folk concept of free will takes a conditional form with respect to determinism. So:

> If the *actual world* is indeterministic, and agents have libertarian powers, then these libertarian powers are what free will is and must be.
> Else, if the actual world is deterministic, and agents have their preferred compatibilist powers, then compatibilist powers are what free will is.

To make things even clearer, this conditional analysis of free will can be organized into a simple two-dimensional diagram (see Figure 1).

		Possible World: I	Possible World: D
Actual World	I	T	F
	D	T	T

'Some agents have free will'

Figure 1. Two-dimensional diagram showing the conditional analysis of free will with respect to determinism, given the sentence 'some agents have free will'.

Here is how to read the two-dimensional table: along the top we see two classes of worlds, indeterministic worlds (I) and deterministic worlds (D). Let's suppose for ease of explication that all indeterministic worlds contain agents with libertarian powers and all deterministic worlds contain agents with compatibilist powers (this assumption can easily be removed with a much more complex diagram). These are 'worlds considered as counterfactual' relative to each other. Down the left-hand side, we see the same two classes of worlds, but here they are not thought of as counterfactual alternatives to each other, where one is actual and the other is an alternative. Instead, they are alternatives about how the actual world itself, for all we know *a priori,* might be.

What we are doing when we read this table, is considering our judgments about whether or not some agents have free will, relative to different contexts (ways things might be, for all we know, only one of which is actual), from the perspective of different indices (ways the actual world might turn out to be). Suppose, then, that the actual world turns out to be indeterministic. From the index of an indeterministic actual world, if we look at counterfactual worlds that are also indeterministic then we will judge that it is true that some agents have free will. This is reflected in the T value in the world at the top left cell of our table. That world is being evaluated from the perspective of an actual indeterministic world (specified on the left of the table). The top right cell contains an F. There, we evaluate what to say about the truth-value of 'some agents have free will' at a deterministic world, from the perspective of an indeterministic

actual world. In that case, since we judge that those deterministic worlds do not contain agents with free will, that sentence comes out as false.

On the other hand, suppose that the actual world turns out to be deterministic. Now consider our judgements about 'some agents have free will' at a deterministic counterfactual world (the cell on the bottom right). Since compatibilist powers are sufficient for free will, we will judge that the sentence is true in that counterfactual world. Furthermore, since having either compatibilist or libertarian powers is sufficient for having free will conditional on the actual world being deterministic, it follows that we will judge that in any worlds with those powers, regardless of whether they are deterministic or not, agents have free will. Hence 'some agents have free will' will be true when evaluated in counterfactual indeterministic worlds, conditional on the actual world being deterministic. This is reflected in the bottom left cell of the table.

Let's tie this back to the empirical results. When a vignette is taken to describe the actual world, we should expect that if people deploy a conditional concept, they will judge that agents are free in the deterministic world considered as actual, and will judge that agents are unfree in the counterfactual deterministic world. People are inclined to judge that people in the counterfactual deterministic world are unfree, because people in fact believe that the actual world is indeterministic and so think, unless told otherwise, that indeterminism is a necessary condition for free will.[7] So, far so good; but this evidence is only consistent with the folk having a conditional concept of free will with respect to determinism. The reason these results do not show that people in fact possess a conditional concept of free will is because we do not have data and responses to all the conditions necessary to determine whether or not there is a conditional concept.

Recently, Latham (2019) tested more directly whether or not the folk concept of free will is a conditional one with respect to determinism. They noted that the conditional account makes two key predictions regarding people's free will judgments to various conditions, which they called the *weak* and *strong signal for conditionality*. The weak signal for conditionality is what Roskies and Nichols (2008) identified might be present in their data. Given that people tend to believe the actual world is

[7] As a descriptive matter of fact, the overwhelming majority of ordinary people think that the actual world is indeterministic. For example, Nichols and Knobe (2007) found over 90% of participants chose the vignette describing an indeterministic universe, not a deterministic universe, as being most like the actual world. Similarly, Latham (2019) found 81.6% of participants selected the indeterministic universe as being most like the actual universe.

indeterministic, if they possess a conditional concept and are asked to evaluate the actual deterministic world, they should be expected to respond that there is free will in such a world. That's because according to the conditional concept, indeterministic and libertarian powers are only necessary for free will if they obtain actually. The strong signal for conditionality was more novel. For the minority of people who believe the actual world is deterministic, if they possess a conditional concept and are asked to evaluate a counterfactual deterministic world from the perspective of an actual indeterministic world, they should be expected to respond that there is no free will in that world. That's because according to the conditional concept, indeterminism and libertarian powers are necessary for free will if the actual world is indeterministic.

Latham (2019) found that people who believe the actual world is indeterministic respond that there is free will in an indeterministic actual world and a counterfactual indeterministic world from the perspective of a deterministic actual world. Further, they respond that there is no free will in a counterfactual deterministic world. Interestingly though, people who believe the actual world is indeterministic are unsure whether or not there is free will in the deterministic actual world (the weak signal for conditionality). People who believe the actual world is deterministic respond that there is free will in the deterministic actual world, the indeterministic actual world, and counterfactual indeterministic actual world from the perspective of a deterministic actual world. Again, interestingly, people who believe the actual world is deterministic are unsure whether or not there is free will in the counterfactual deterministic world, from the perspective of an indeterministic world (the strong signal for conditionality).

While people don't straightforwardly respond in a manner predicted by the conditional concept, they do respond in a manner that supports the idea that we possess a conditional concept with respect to determinism. That's because I don't think it is mere coincidence that people who believe the actual world is indeterministic are unsure how to respond to an actual deterministic world. Nor do I think it's a coincidence that people who believe the actual world is deterministic are unsure how to respond to a counterfactual deterministic world from the perspective of an actual indeterministic world. Both these conditions are correctly identified as being important with respect to people's concept of free will once it has been identified that our concept of free will might be a conditional concept with respect to determinism.

Why are people unsure how to respond in conditions associated with the weak and strong signal for conditionality? Let's start with the weak signal

for conditionality. Imagine someone believes the actual world is indeterministic and is then asked to evaluate whether there is free will in the actual deterministic world. It's extremely unlikely that people change their beliefs about the actual world in order to perform such evaluations. Instead, what people most likely do is simulate how they would respond if they counterfactually believed the actual world is deterministic. Importantly, this cognitive process does not mask the effects of what people actually believe, which is what explains why people are unsure about how to respond. If someone has a conditional concept and believes the actual world is indeterministic, then they should also think that indeterminism and libertarian powers are necessary for free will. So according to their actual belief there is no free will in the deterministic actual world. But if they succeed in simulating what they would think if they counterfactually believed the actual world is deterministic, then they should also think compatibilist powers are sufficient for free will. So according to their simulated counterfactual belief there is free will in the actual deterministic world. Thus, there is a response conflict between their responses generated in accordance with their actual belief, and their simulated counterfactual belief.

This also explains why we observe that people who believe the actual world is deterministic are unsure how to respond in the condition associated with the strong signal for conditionality. Imagine now someone who believes the actual world is deterministic and is asked to evaluate whether there is free will in a counterfactual deterministic world from the perspective of an indeterministic actual world. If that person has a conditional concept with respect to determinism, then they should also think that compatibilist powers are sufficient for free will. So according to their actual belief there is free will in the counterfactual deterministic world. But if they succeed in simulating what they would think if they counterfactually believed the actual world is indeterministic, then they should no longer think that compatibilist powers are sufficient for free will. Instead they should think that indeterminism and libertarian powers are necessary for free will. So according to their simulated counterfactual belief there is no free will in the counterfactual deterministic world. As a result, there is a conflict between free will responses that are generated in accordance with someone's actual and simulated counterfactual beliefs.

3.2. Determinism and the Conceptual Impossibility Thesis

If the folk concept of free will is a conditional concept with respect to determinism, then the conceptual impossibility thesis too, at least with respect to determinism, is correct. That's because no matter how things turn out actually to be—with respect to the world being deterministic or

not—if we possess that concept, we will judge that we possess free will. Once you hold fixed the compatibilist powers and libertarian powers in all these worlds, all worlds considered as ways things might actually be contain agents with free will. So even if determinism is actually true, and we only possess compatibilist powers, we will judge that we are free. On the other hand, if indeterminism is actually true, and we possess libertarian powers, we will judge that we are free, and that indeterminism and libertarian powers are necessary for free will.

This means there is something we could discover, if the conditional story is correct, which would make us think that indeterminism and libertarian powers are necessary. But that doesn't mean that the conditional concept of free will is inconsistent with the conceptual impossibility thesis, because there is nothing we could discover about how things are *actually* that would make us judge that actually there's no free will. Remember, we're holding fixed here that there are actually either compatibilist or libertarian powers. So, my claim is just that nothing we could discover about determinism would lead us to judge that we are unfree. As I suggested earlier, I think the conceptual impossibility thesis generalizes beyond determinism, but I have no empirical data that can support that contention here. Still in the next section (§4) I will give some good reasons why I think we should expect this.

So, with regard to the world being deterministic or not, free will is compatible with anything that we could discover about how things actually are. But if that's right then how did we become convinced that the folk concept of free will is an incompatibilist one? The conditional analysis offers up a ready explanation. If people think that the actual world is indeterministic and contains agents with libertarian powers, then they will judge not only that we are free, but also that deterministic possible worlds containing only agents with compatibilist powers lack free will (see footnote 7). So to the extent people are confident that actually, the world is indeterministic and there are libertarian powers, they should be expected to deny that compatibilist powers are sufficient for free will. From the perspective of a world where indeterminism is true, some, but not all, counterfactual worlds will contain agents with free will.

Of course, in most of the free will literature the distinction between judging of the actual world that it is deterministic and that indeterminism is a necessary condition for free will, and judging of the actual world that it is indeterministic, and that indeterminism is a necessary condition for free

will, is not made.[8] What actual philosophers of free will, embedded and entrenched in their philosophical views, would judge when this distinction is drawn is not something about which we have empirical data. Nevertheless, I think this distinction makes a difference to the judgments of ordinary agents.

4. Conditionality and the Conceptual Impossibility Thesis

In the previous section I provided evidence that the folk concept of free will is conditional with respect to determinism. This results in our judgments about whether or not *we* typically possess free will being insensitive to whether determinism is actually true. Instead, the truth or otherwise of determinism only affects our counterfactual judgments about whether agents in other worlds have free will. One way to think of the conceptual impossibility thesis is as a generalization of this.

So far, I have talked about whether determinism is true or false *simpliciter*. But it's important to also consider potential defeaters of free will (of which determinism is just one) in another way: the local way. How might we react if we were to learn that it is *sometimes*, *somewhere* true. While in fact in the case of determinism it is plausible that it's either globally true, or else false, when we generalize from determinism to other factors people might have thought important for free will, this may not be so.

Our judgments about whether we typically have free will are *insensitive* to various apparent defeaters to our free will being true *in general*. Imagine for the moment your favorite free will defeater *X*. If there is no global *X*, then having *X* rules out counterfactual populations from being free (and perhaps niche local populations as well). But if in fact *X* is generally actually true, then it doesn't affect our judgments about counterfactual populations. The presence or absence of *X* does *not* affect out judgements about whether actually *we* are free at all.

Let's work through a couple of examples. Imagine how the account I am offering might deal with another important challenge to free will. If what some brain scientists think is correct, then conscious psychological states do not perform the role we suppose they do for our actions (e.g., Libet et

[8] To the best of my knowledge Peter Van Inwagen (1983) is the only theorist who appears to identify this distinction and thinks that the actual world is indeterministic and that this indeterminism is necessary for free will. In the very last paragraph of his book *An Essay on Free Will*, he writes "…it is conceivable that science will one day present us with compelling reasons for believing in determinism. Then, and only then, I think should we become compatibilists." (p. 223)

al. 1983; Soon et al. 2008). Instead, conscious psychological states, and the actions that we suppose they cause, are both caused by an unconscious common cause. Let's call worlds where all actions are like these brain scientists think: *Libet worlds*. On the account I have been describing, if the actual world is one where conscious processes are causally involved in typical decisions, then we might think that is necessary for free will. But if actually they are not—if our world is a Libet world—then we will say that so long as the typical neural common cause of the action and its accompanying conscious state is in place, then the resultant action is free.

We can also imagine an even more extreme case (even by the lights of the free will literature). Imagine everyone's actions everywhere are being controlled by an alien species called Dromes. These Dromes have total control over both our conscious and unconscious psychological states, and thus our actions as well. For ease of explication, let's call worlds where all actions are controlled by Dromes: *Drome worlds*. On the account I have been describing, if the actual world is a Drome world, then we would still have free will, since free will is just whatever we are tracking that that distinguishes the cases we ordinarily judge to be free and the cases we ordinarily judge to be unfree. But if the actual world is not a Drome world, as is commonly supposed, then only those actions that are not the result of Drome control will be free, and necessarily so.[9]

Of course, we can be almost certain that free will error theory would be true of our metaphysical concept of free will if the actual world is either a Libet or Drome World. Still, despite there being no deep metaphysical difference between the cases we judge to be free and unfree, I think that we can be expected to want our actions to be like the ones that we ordinarily think of as free. Even if, with respect to some particular metaphysical matter of fact, there is no difference between these actions, there are relevant differences that our social concept of free will tracks with our talk of free and unfree action. It also seems that we can be expected to normatively evaluate actions, and to do that we need to distinguish actions that are performed while being, (what we might have ordinarily of thought of as), coerced and manipulated, from those that are not, and actions performed while being in the grips of, (what we might have ordinarily of thought was), a psychological or physiological illness, from those that are not, and so on. Regardless of whether all our actions are the result of unconscious processes, Dromes, *mutatis mutandis* for all these kinds of

[9] You might think that it's consistent with us making the discovery that we have no free will that our free will practices would persist, albeit as a useful fiction. However, on the view that I am advancing here, if the free will practices are what is in common between paradigm cases, then free will just is realism about the practices, and so we do have free will.

cases, we still care that our actions be like the ones that we would ordinarily think of as being free. If that's right, then the conceptual impossibility thesis is true of our social concept of free will.

5. Conclusion: The Conceptual Impossibility of Free Will Error Theory

In this paper I have argued for the conceptual impossibility of free will error theory - the conceptual impossibility thesis. There are two apparent concepts of free will: a metaphysical concept that tracks some metaphysical feature of our actions, and a social concept that tracks relevant differences between actions we ordinarily judge to be free and unfree. The weak conceptual impossibility thesis is that while free will error theory might be true on the metaphysical concept, there will be free will regardless, on the social concept. That's because our social concept of free will cannot do the job it's supposed to, and that concept fail to be satisfied. So, the conceptual impossibility thesis is true of that concept. The strong conceptual impossibility thesis is that while both concepts exist, only the social concept matters, and so the conceptual impossibility thesis is true of the only concept of free will we care about.

The conceptual impossibility thesis not only makes good sense of our practices—that we continue to hold people responsible for some actions, and not others, regardless of whether we think that our world is deterministic, and regardless of whether we think that certain neuroscientific findings hold—and it helps us make sense of some inconsistent findings in the experimental philosophy literature examining our concept of free will. This, jointly, gives us some reason to think that the conceptual impossibility thesis is correct, and that there is no way the actual world could be such that we judge that we do not have free will: on at least *one* of our concepts.

Acknowledgements

I am grateful to David Braddon-Mitchell, Kristie Miller, Michael Duncan, James Norton, two anonymous referees, and the metaphysics group at the University of Sydney for their useful feedback on the earlier versions of this manuscript. Thanks to the Ngāi Tai Ki Tāmaki Tribal Trust for their support.

REFERENCES

Björnsson, G. 2014. Incompatibilism and "Bypassed" Agency. In *Surrounding Free Will*, ed. A. R. Mele. Oxford: Oxford University Press.

Braddon-Mitchell, D. 2003. Qualia and analytical conditionals. *The Journal of Philosophy*. 100: 111–135.

Braddon-Mitchell, D., and K. Miller. 2019. Quantum gravity, timelessness, and the contents of thought. *Philosophical Studies*, 176: 1807–1829

Chalmers, D. J. 2005. The Matrix as Metaphysics. In *Philosophers Explore the Matrix*, ed. C. Grau. Oxford: Oxford University Press.

Churchland, P. M. 1981. Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, 78: 67–90.

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

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

Deery, O. 2019. Free actions as a natural kind, *Synthese*. DOI: 10.1007/s11229-018-02068-7

Dennett, D. C. 1984. *Elbow Room: The Varieties of Free Will worth Wanting.* MIT Press.

Dennett, D. C. 2013. Please Don't Feed the Bugbears. In *The Philosophy of Free Will: Essential Readings From the Contemporary Debates*, eds. P. Russell and O. Deery. Oxford: Oxford University Press.

Ekstrom, L. 2002. Libertarianism and Frankfurt-style cases. In *The Oxford Handbook of Free Will, 2nd edition*, ed. R. Kane. Oxford: Oxford University Press.

Flew, A. 1955. Divine Omnipotence and Human Freedom. In *New Essays in Philosophical Theology*, eds. A. Flew and A. McIntyre. London: SCM Press.

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

Kane, R. 2005. *A Contemporary Introduction to Free Will*. New York: Oxford University Press.

Kant, I. 1781/1787. *Critique of pure reason.* P. Guyer and A. Wood (eds. and trans.), Cambridge: Cambridge University Press, 1997.

Latham, A. J. 2019. *Indirect Compatibilism*. Dissertation. http://hdl.handle.net/2123/20440

Libet, B., C. A., Gleason, E. W. Wright, and D. K. Pearl. 1983. Time of conscious intention to act in relation to onset of cerebral activity

(readiness-potential). The unconscious initiation of a freely voluntary act. *Brain*, 106: 623–42.

Lockie, R. 2018. *Free Will and Epistemology: A Defence of the Transcendental Argument for Freedom*. London: Bloomsbury Academic.

Nahmias, E. 2011. Intuitions about Free Will, Determinism, and Bypassing. In *The Oxford Handbook of Free Will, 2nd edition*, ed. R. Kane. Oxford: Oxford University Press.

Nahmias, E., S. Morris, T. Nadelhoffer, and J. Turner. 2005. Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology,* 18: 561–584.

Nahmias, E., S. Morris, T. Nadelhoffer, and J. Turner. 2006. Is incompatibilism intuitive? *Philosophy and Phenomenological Research,* 73: 28–53.

Nichols, S. B., and J. Knobe. 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs*. 41: 663-685.

O'Connor, T. 2000. *Persons and Causes: The Metaphysics of Free Will*. Oxford: Oxford University Press.

Pereboom, D. 2001. *Living Without Free Will*. Cambridge University Press.

Pink, T. 2004. *Free Will: A Very Short Introduction*. Oxford: Oxford University Press.

Roskies, A. L., and S. B. Nichols. 2008. Bring moral responsibility down to earth. *Journal of Philosophy*, 105: 371–388.

Soon, C. S., M. Brass, H. J. Heinze, and J. D. Haynes. 2008. Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11: 543–545.

Strawson, G. 1986. *Freedom and Belief*. Oxford: Clarendon Press.

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

van Inwagen, P. 1993. *Metaphysics*. Boulder, Co.: Westview Press.

EuJAP | Vol. 15 | No. 2 | 2019
UDK: 123.1 124.6
https://doi.org/10.31820/ejap.15.2.6

CAN SELF-DETERMINED ACTIONS BE PREDICTABLE?

AMIT PUNDIK
Tel Aviv University

Original scientific article – Received: 30/05/2019 Accepted: 15/09/2019

ABSTRACT

This paper examines Lockie's theory of libertarian self-determinism in light of the question of prediction: "Can we know (or justifiably believe) how an agent will act, or is likely to act, freely?" I argue that, when Lockie's theory is taken to its full logical extent, free actions cannot be predicted to any degree of accuracy because, even if they have probabilities, these cannot be known. However, I suggest that this implication of his theory is actually advantageous, because it is able to explain and justify an important feature of the practices we use to determine whether someone has acted culpably: our hostility to the use of predictive evidence.

Keywords*: Free will, causation, objective probability, determinism, criminal responsibility, Dennett, prediction, Lockie*

1. Introduction

Some philosophers arrive at the free will question from an ontological starting point (for example, "What kind of freedom exists, if any?" or "What are its conditions?"). Others arrive from an ethical starting point (for example, "How should I treat myself or others when we fail to do what we ought to?"). By contrast, Lockie takes a refreshing epistemic stance — based on the forceful transcendental argument for libertarianism that his book presents — and questions how epistemic norms affect the arguments we can use to support metaphysical claims about free will. Here, I would like to focus on the specific account of libertarianism that he proposes: his theory of self-determination (outlined mainly in Chapter 9). I will examine

his account in light of a different epistemic question, that of prediction: "Can we know (or justifiably believe) how an agent will act, or is likely to act, *freely*?" I would argue that, when Lockie's self-determinism is taken to its full logical extent, free actions cannot be predicted to any degree of accuracy on the basis of anything other than previous free actions, because even if they have probabilities, these cannot be known. While Lockie himself seems to accept the view that free actions may be predictable, I argue that this view cannot be accommodated with other parts of his theory, hence he needs to choose between freedom and predictability. Furthermore, I would like to suggest that this implication of his theory, that free actions cannot be predicted, is actually advantageous: it is able to explain an important feature of the practices we use to determine whether someone has acted culpably—particularly, though not only, in criminal trials.

I start by arguing that, to be epistemically warranted, predictions need to rely on causal generalisations. I then turn to Lockie's self-determinism and examine whether the agent's character traits, reasons, and objective probabilities, or maybe even the agent as a whole, may be used to anchor such causal generalisations. Lastly, I briefly explain the hostility of Common Law to predictive evidence and suggest that libertarian theories that renounce the idea that free actions have discoverable objective probabilities are able to account for this hostility.

2. Why Predictions Require Causal Generalisations

Inferences from a known to an unknown empirical fact involve a generalisation about types. Schauer, for example, holds that "the avoidance of generalizations is, with few or no qualifications, simply not possible at all" (Schauer 2003, 101). In some cases, the reference to the generalisation is made explicitly. For example, inferring that Socrates is mortal from our knowledge that human beings are mortal refers explicitly to a generalisation about human beings as a type. However, in many cases the generalisation is implicit in the inference. Consider, for example, an inference from the fact that a person reacted allergically to a certain cat to the fact that this individual is likely to react allergically to that same cat in future. This knowledge implies one or more generalisations that could serve as the basis for the inference (for example, the type of person who once reacted allergically to cats is likely to continue to react allergically). The important point is that drawing an inference from one empirical fact to another presupposes a generalisation about types of fact that connects the fact from which the inference begins to the fact with which the inference ends. Without this presupposition, the inference is invalid

because it remains unclear what licenses the move from the first fact to the second.

I contend that inferences from a known to an unknown empirical fact require a causal generalisation — that is, a generalisation that reflects a causal connection between the type of fact from which the inference begins and the type of fact the inference seeks to establish. If an inference is based on a non-causal generalisation, a mere correlation, it is unlicensed and thus invalid (this claim is part of the Common Cause Principle, see Reichenbach 1999, 157-166; Arntzenius 1992). The causal relation can operate either directly or through a common cause. For instance, inferring that a smoker is likelier to contract cancer than a non-smoker is based on a causal generalisation that smoking is a cause of (lung) cancer. By contrast, inferring that a Coca-Cola drinker is likelier to contract cancer than a non-drinker involves a causal generalisation that reflects a common cause. It is living in a hot country that is the common cause of both Coca-Cola drinking and (skin) cancer. I do not argue that, for the inference to be valid, it is necessary to specify the (direct or indirect) causal generalisation; I only argue that the existence of such a causal generalisation has to be presupposed.

Consider the opposite stance, according to which a mere correlation between two types of fact can suffice to infer an unknown from a known fact, without making any commitment about the existence or kind of causal connection between these types of fact. Such a stance would still require that the generalisation on which a valid inference is based satisfy certain conditions or standards, such as statistical significance. The difficulty with such a stance is that it renders the rejection of spurious correlations more difficult. Spurious correlations are those that do not reflect any actual connection (be they causal or not) between the two types of fact. Consider the correlation between the number of people who drowned by falling into a swimming pool during a given period of years and the number of films in which Nicolas Cage appeared, in that same period (Vigen 2015). The lack of any actual connection between these facts means that this spurious correlation does not hold outside the group of initially-observed cases. It would hence be a mistake to infer anything about the number of people who drowned from the number of Nicholas Cage films (or vice versa) in a year that is not included in the group of years within which the correlation was identified. Drawing any inference from a spurious correlation to an unobserved case is therefore unlicensed and misleading, whatever the purpose of the inquiry is (be it to obtain knowledge, provide an explanation, or make a prediction about unobserved cases). Identifying a reliable process to ensure that a given correlation is not spurious is therefore essential, because spurious correlations are so widespread —

indeed, they are bound to be ever-present. Since each specific case consists of innumerable details (most of which are, of course, unimportant), one could sift through a vast number of facts until one finds a group in which the identified fact correlates with the fact that one seeks to establish. For example, one might find a correlation between a certain type of action and the second (or third) letter of the person's great-aunt's surname.

If one accepts that inferences require causal generalisations, one can apply methods to distinguish between causal and non-causal connection to identify which generalisations are spurious (for the various sophisticated methods that have been proposed, such as the Markov Condition and Bayesian Nets, see Williamson 2005). However, if one denies that inferences require causal generalisations, one ought to find how to distinguish between informative and spurious correlations. Note that mere statistical significance will not do, because testing sufficiently large numbers of variables using sufficiently large databases would eventually generate statistically significant (yet spurious) generalisations. One might wish that such absurd, albeit statistically-significant, correlations simply did not exist. But this wish relies on the assumption that statistically-significant correlations need to "make sense" — that is, that it would be possible to explain why this correlation holds; and what would such an explanation be, if not causal or causal-like?

One might challenge this argument using counterexamples in which an inference from a known to an unknown fact is made without presupposing a causal connection between the types of fact. For example, if there are ten balls in a jar, of which nine are blue, it might be possible to infer that the probability of a randomly-chosen ball's being blue is 90%, without presupposing any causal connection between "being in that jar" and "being blue".

However, even if not all factual inferences require a causal connection to be presupposed, the inferences drawn in legal fact-finding almost always do. Denying an underlying causal connection is easier when the generalisation is extracted from a group of cases to which the specific case at hand belongs. It is important to note that the randomly-chosen ball is, itself, one of the ten balls in the jar. It might thus be possible to draw some inferences about it without presupposing anything about the relation between the types of fact. While such inferences raise a set of difficult problems (Hájek 2009), these differ in kind from those involved in drawing inferences from generalisations that do not include the case at hand. To infer the probability that a randomly-chosen ball will be blue from the proportion of blue balls in another jar, it is necessary to presuppose that there is some substantial relation between "being in a jar" and "being blue".

And, again, if this substantial relation is not causal or causal-like, what else could it be?

3. The Predictability of Self-determined Actions

Instructive hints on the predictability of self-determined actions can be found in Lockie's discussion of Dennett's character-based example of a person who is unable to torture an innocent for $10 (Lockie 2018, 216). Let us ignore Milgram's experiments and assume, with Dennett, that this is, indeed, a paradigmatic example of an action (or omission) that is determined. Dennett uses this example to argue that the fact that the person's actions are determined (in this example, by his moral nature) does not undermine freedom. Dennett seems to rely on an intuition that this person acts freely when he does not torture. I do not share this intuition. My view is that the person's omission, if so determined, is *unfree* because they had no reason to torture (and it therefore seems to me that they deserve no praise for this omission — though I will not pursue this point here). Lockie, however, agrees with Dennett that the person's omission to torture for $10 is free. He explains: "I may be unable to deviate from a path that I, my moral nature, has determined" (*ibid.*). However, he insists that Dennett's example fails to establish compatibilism, because the individual could still be free even if so determined: "Ethical responsibility […] is something that requires freedom from determination by the Big Bang and laws of nature precisely in order to preserve the possibility of self-attributable axiological determination: of the agent determining acts in accordance with his moral nature and responsiveness to moral (and other) reasons" (*ibid.*).

I would argue that Lockie's agreement with Dennett cannot be settled with the rest of his theory, hence Lockie needs to accept that the person's omission to torture is unfree. More generally, I would suggest that Lockie has no theoretical resources to explain how actions may be both free and predictable based on anything other than the agent's own previous free actions. I would seek to establish this claim by discussing a related question: while Lockie discusses the matter of whether this omission could be both free *and* determined (by the person's moral nature), I would like to question whether this omission could be both free and *predictable*. I assume that, if any human actions and omissions are predictable, a person's omission to torture for $10 must be one of them, even if nothing is known about the previous actions of that person. As I have already argued, this prediction must rely on a causal generalisation. Consequently, the question revolves around what the *relatum* of the causal connection that this generalisation reflects might be.

The first possibility is that the person's "moral nature" is the relatum, causing the agent to refrain from torturing an innocent. Moral nature itself could be the relatum, or it could consist of some propensities, traits, and so on, one of which determines the agent's omission. Whatever the exact relatum is, under this option, moral nature (or one of its components) is *ontologically distinct* from the agent. Such a view could easily explain the predictability of this omission: to have a certain moral nature either implies or consists of the *predictable tendency* of the agent to act in a certain way. For example, to be a kind-natured person is to have a higher likelihood of performing kind actions (compared to another person who is of an unkind nature). However, Lockie rightly rejects moral nature as the relatum that causes the agent's omission. He asks: "whatever the conative part was that determined your choice, was this in turn determined by natural law or by chance?", to which he answers: "persons, not their parts, determine choices" (*ibid.*, 196). Lockie makes this move to fend off Hobbes' regress objection: if some part of the person determined the choice, what determined that part? Consequently, Lockie objects to "an ontologically real, prior and separable item in the chooser called an act of will — historically: a 'desire', or sometimes 'volition'" (*ibid.*) As a result, according to Lockie, neither moral nature, nor any other part of the agent's character, can be the causal relatum on which predictions would be based.

The second possibility is that the predictability of the refusal to torture is based on the person's *reasons*. Perhaps these could explain why it is so predictable that they would refuse to torture an innocent for $10. While their reason for doing so is weak ($10), they have plenty of forceful reasons to avoid torturing. But Lockie seems to take the view that reasons are not causes: "For [reasons] to play a role, they don't associatively cause action and cognition. They enter a *mind* and become active in interaction in that mind. The mind (the agent, the person, the self) decides — in active assimilation, accommodation and equilibration of *that agent's* reasons" (*ibid.*, 207). Yet, if reasons have no independent causal power, they cannot constitute the relatum in the causal generalisations needed to make our predictions warranted.

Moving to the third possibility, perhaps the predictability of self-determined actions could be rooted in objective probabilities. One helpful way to understand objective probabilities, for our purposes, is as *free-standing ontological entities* (otherwise, it is unclear how they could be the relatum in the causal generalisations on which warranted predictions are based). Consider the predictability of the radioactive decay of a certain unstable atom. Let us assume that this atom, X, has a probability, Y, of decaying in the next second. Let us assume further that this is not a

subjective probability — even if we know everything that could possibly be known about this atom and the laws of nature, we would still be unable to predict *with certainty* whether this atom will decay in the next second. However, we can still predict that X will decay *with a certain probability*. After we observe enough cases of such atoms, we could generalise that the probability that X will decay in the next second is Y. The closer our prediction gets to the objective probability, the more accurate it is. This prediction is warranted because it is based on a causal generalisation in which objective probabilities are either the relatum or part of it: the objective probability is ontologically distinct from the other parts of the causal relations. It either causes the effect directly or it allows other potential causes to bring about the effect (or prevents them from doing so).

If free actions have such ontologically-distinct objective probabilities, they could warrant our predictions. If there is an objective probability that the agent will torture an innocent for $10, and that probability is zero or close to zero, it could anchor our prediction. Under this view, predicting that the person is not going to torture is warranted, because it is based on a causal generalisation in which objective probabilities are part of the causal relatum that, together with the agent, determines the action. The person's inability to torture is basically a prediction that reflects this close-to-zero objective probability that they *would* torture.

However, if objective probabilities are such *free-standing ontological entities*, they cannot help make the agent's action self-determined. On the contrary, they would get in the way. Based on James, Lockie distinguishes between positive and negative chances. A positive chance is "a true generator of randomness in the world", which is inserted "into, or over, or at, the origin of our acts" (*ibid.*, 197). Such a chance is "destructive of freedom and responsibility" because "responsibility for our actions as remained to us would be just that degree of determination of our actions as was robust enough to survive this chaos, this noise, these gremlins. The more noise, the less *we* would determine action — the less we could be said to act at all" (*ibid.*, 197-8). By contrast, a negative chance "is simply an absence of determination by any positive force *external to the agent himself*" (*ibid.*, 198, original emphasis). So Lockie rules out the type of chance required for rooting our predictions in objective probabilities, *at least when they are understood as free-standing ontological entities.*

Let us take stock: predicting the agent's free actions requires a causal generalisation, but neither the agent's character traits and propensities, nor their reasons, nor even ontologically-separable objective probabilities could be used as the relatum. The fourth and last possibility I can think of is that the relatum consists of the *agent himself*, as a whole. Being an

agent–causalist myself, I find this view very plausible — but the question is how such a view would account for predictions. Can such an agent have objective probabilities that are *integral* to them, making them who they are, rather than being "ontologically real, prior and separable items"? I would argue that, even if such objective probabilities were somehow possible, it would not matter for any practical purpose; in particular, it would not warrant our predictions. This is because such objective probabilities, even if they do exist, *cannot be known*.

When distinguishing the agent's character from their reasons, Lockie emphasises that "[t]he person's character is, in a significant sense, *ontologically unique*, prior and fundamental" (*ibid.*, 207). I take it that, when Lockie refers to "the person's character" he means *the agent as a whole*, in an attempt to distinguish the agent from their reasons: "… what it is to have the character of the one isn't just to be built up out of ('bundled out of') different, and different-strength, 'reasons' — it is to be ontologically different; it is to be a different person" (*ibid.*, 206). It is little wonder that Lockie emphasises uniqueness: if an agent is self-determined, and not subject to any natural law, what would be the basis for assuming that one self-determined agent will determine themselves similarly to another, if there is no external law to govern their conduct? But *if the agent, person, or character is unique, how could we predict what the agent will do freely, if we cannot draw any inference from observations about other similar people?*

One might respond that we could still predict that the person is not going to torture an innocent for $10, based on his previous actions. However, this response does not help to reconcile Lockie's agreement with Dennett with the rest of his theory, because Dennett's example seems to work even in cases in which the prediction that the person will refuse to torture for $10 is not based on their previous actions. I can step into my classroom on the first day of the academic year, knowing virtually nothing about the 150-or-so students there, and yet predict that they would not torture an innocent for $10. If predictions are warranted only based on my knowledge of previous actions, then it is unclear how this prediction is warranted. More generally, setting aside cases in which a free action is predicted based on the agent's previous actions, Lockie has no resources to explain how actions can be both free *and* predictable.

4. Is Unpredictability of Free Actions a Disadvantage for Lockie's Theory?

A constitutive feature of libertarian theories of free will is the claim that, if the agent's action were (fully) determined by antecedent causal factors outside their control, they would be neither free to do, nor culpable for doing, what they did. Yet, libertarians tend to accept the view that the agent's *free* actions have *objective probabilities* (van Inwagen 2000, 14-18; O'Connor 2000, 97; 2009, 197), and that position is rarely challenged (for exceptions, see Vicens 2016; Sela 2017). If Lockie's self-determined actions cannot be predictable, as I argued in the previous section, this implication of his theory might be viewed as a serious problem, even by those who are sympathetic to his libertarian inclinations. By contrast, my view is that accepting unpredictability as a necessary condition of free will may assist libertarian theories to overcome some of the common objections levelled against them.[1] In the remainder of this paper, I would like to suggest that accepting this implication of Lockie's self-determinism has the advantage of being able to explain an important feature of the practices used to determine whether someone has acted culpably — particularly, but not exclusively, in criminal trials.

While I believe that the following analysis is applicable more widely, to legal and non-legal practices of determining culpability alike, I focus here on the former because they include explicit and well-specified rules. I take legal practices in criminal trials to offer the most suitable case study because criminal punishment is clearly constrained by culpability, at least if criminal law seeks to avoid punishing those who are not culpable for their actions. This constraint does not imply retributivism — namely, that punishment is inflicted because it is deserved. Instead, any theory of punishment that considers culpability to be a *constraint* on other legitimate goals of punishment should refrain from knowingly convicting the innocent.[2] Hence, criminal proceedings constitute the clearest context in which culpability is attributed. I also assume, like many theorists of free will, that acting freely or with some kind of control is a necessary condition of culpability. While some might hold that our practices of attributing culpability do not require us to settle the metaphysical problem of free will (Strawson 1962), I share the position that the distinction between justified and unjustified attribution of culpability — which any theory of culpability

[1] For example, unpredictability may assist libertarians to overcome van Inwagen's rollback argument. See Bernáth and Tőzsér (2019).

[2] One notable example of such a theory is Hart's mixed theory, which accepts the retributivist constraint ("only those who have broken the law—and voluntarily broken it—may be punished") while rejecting retributivism as the "General Justifying Aim of the system" (Hart 2008, 9).

seems to need — is likely to rely on (or bring through the back door) notions very similar to "freedom" and "control" (Tadros 2005, 69).

The scope of my discussion is restricted in one important respect. Some culpable actions may cause the agent to perform further actions that may be both predictable and culpable (getting drunk voluntarily and then driving dangerously). The agent's culpability for the latter seems to be *derived* from their culpability for the former. When, how, and why culpability for one action is derived from another are complicated issues to address, and it is particularly questionable whether the agent's culpability goes beyond their culpability for the first action. Be that as it may, such derivatively-culpable actions are outside the scope of this paper. I will therefore not discuss here evidence of planning, preparation, and motive (because, in such cases, the evidence may be probative of the alleged crime by establishing an earlier free decision that caused both the creation of the predictive evidence and the later commission of the crime).

When determining, in criminal proceedings, whether an individual performed a certain culpable action, predictive evidence is often ignored.[3] Most apparently, and with only a few exceptions, base-rates are excluded (Koehler 2002). Using such evidence in court also seems *intuitively problematic*. For example, using the high rate of crimes involving illegal firearms in a certain neighbourhood to support the conviction of an individual resident in a crime involving an illegal firearm (henceforth, the "crime-rates scenario") seems highly objectionable. The objection to base-rates is not only aimed at the sufficiency of such evidence (on the grounds that "crime-rates are insufficient on their own to prove that the individual is guilty"). The objection also requires that such evidence not be used at all in determining the individual's guilt: that crime-rates be *inadmissible* in criminal proceedings.[4] The hostility of criminal fact-finding toward predictive evidence is also apparent in the deeply-rooted suspicion of bad character and previous convictions.[5]

[3] I rely on Uviller's distinction between trace and predictive evidence: the former results from a past event that leaves some traces in the present (e.g. eyewitnesses, fingerprints), while the latter "looks forward from an established event or trait to predict the likely repetition of its occurrence" (Uviller 1982, 847).

[4] This intuitive objection to admissibility distinguishes this example from the lottery and preface paradoxes in epistemology and the gate-crasher and prisoners paradoxes in legal theory. I have argued elsewhere that the latter are confusing and unhelpful; see Pundik (2017, 192-193).

[5] "English law's suspicion of bad character and extraneous misconduct evidence has been cultivated for many centuries. It is deeply embedded in English judicial culture and institutions, and has frequently been actively propounded and celebrated" (Roberts and Zuckerman 2010, 586).

Legal scholarship contains various accounts that seek to justify the exclusion of such predictive evidence. The first kind of strategy, which has received the most scholarly attention, aims to identify an epistemic deficiency in the inference made from predictive evidence to the specific case. The inference is lacking: in weight (Cohen 1977, 74); appropriate causal connection (Thomson 1986); case-specificity (Stein 2005, 64-106); ability to provide the best explanation (Dant 1988; Allen and Pardo 2008); immunity to the problem of the reference class (Allen and Pardo 2007); or sensitivity to the truth (Enoch et al. 2012).[6] I am unconvinced by these epistemic accounts, because I think that not only does each one suffer from its own problems (Pundik 2008a), but they all share some common deficiencies (Pundik 2011; see also Schoeman 1987 and Redmayne 2008). For example, why should the very same inference that is condemned as epistemically objectionable nevertheless be good enough for prediction purposes? If the inference suffers from some epistemic deficiency, this deficiency arises not only in the context of conviction but also in that of prediction.

The second kind of strategy seeks to identify something in the *legal context* that makes some uses of predictive evidence objectionable, such as the rituality of the legal process (Tribe 1971), the over-transparency of standards of proof (Nesson 1985), equality between litigants (Stein 2005, 105), and the individuality and autonomy of the litigant against whom the evidence is used (Wasserman 1992; Zuckerman 1986). Proponents of this type of account share the view that, even if such evidence may be useful in other contexts (science, policymaking, and so on), its use in legal fact-finding conflicts with fundamental values of the legal system. I believe that, while there are specific problems with each of these accounts,[7] they capture something significant about predictive evidence because their strategy easily explains why the appropriateness of using this evidence depends fundamentally on the purpose for which it is used.

In previous work,[8] I have suggested a contextualist account that is based on culpability. According to this "culpability account", some types of

[6] The reference is to the epistemic explanation appearing in the first part of their paper, although, in the second, they argue that epistemic considerations do not suffice to exclude predictive evidence, and later propose an alternative account based on primary incentives.
[7] See Schoeman (1987). For criticism of Nesson and Tribe's accounts, see Shaviro (1989). For criticism of Wasserman's, see Pundik (2008b). For criticism of Stein's, see, e.g., Pundik (2006).
[8] This section rehearses the argument I made in Pundik (2017). Given the complexity of the issues involved (causation, free will, and so on), I chose to repeat the argument itself in full but to remove some of the more nuanced qualifications. Readers who are not familiar with that paper and are left with some concerns about the claims made might find replies in there.

generalisation about human conduct presuppose that the individual's conduct was determined by a certain *causal* factor that rendered their conduct unfree. By contrast, in the context of attributing culpability, it is necessary to presuppose the exact opposite: that the accused was free to determine their own conduct. Using these types of generalisation to determine culpability is objectionable, because it involves *contradicting* presuppositions about the individual's conduct.

In Section 2, I argued that inferences about human conduct require reliance on causal generalisations. But, even if they do, why can free actions not be proven with such generalisations? Starting with a simple example, assume that Richard is exposed to radiation of a particular kind, which affects his nervous system, resulting in blotches all over his skin and an irresistible urge to attack everyone around him. Assume further that *every* person exposed to this radiation develops these symptoms. When Richard is admitted to hospital, it seems unproblematic to infer from the blotches that, given the opportunity, he will go berserk and should therefore be restrained. However, inferring from these blotches that a violent action that had taken place before Richard arrived at the hospital was committed by him (rather than by someone else), for the purpose of convicting him of a violent offence, seems intuitively problematic.

According to the culpability account, this inference should not be used for the purpose of determining culpability, because it leads to a contradiction. To infer from Richard's skin blotches that he had acted violently, it is necessary to presuppose a *causal* generalisation: either one caused the other or they both have a common cause. In this example, the radiation caused both Richard's blotches and his violent conduct. However, Richard's acting violently may be culpable only if he acted *freely*. The culpability account is based on a libertarian theory of free will, which holds that people do not act freely when their conduct is determined by antecedent conditions outside their control. Establishing Richard's guilt by inferring from the blotches on his skin that it was he who acted violently is, therefore, contradictory: Richard's conduct is treated as free and unfree at the same time.

Blaming Richard for a violent action, having inferred his conduct from the blotches, is problematic, since such an inference cannot be used without dissolving his culpability. Similarly, if the inference is used to predict that Richard *will* act violently, it is only at the price of implying that his violent conduct will not be culpable. This example also explains why the very same inference seems unproblematic when restraining him in the hospital. While inferring from the blotches that Richard will act violently in the hospital presupposes that his conduct is determined (and hence unfree),

this leads to no contradiction because, in the medical context, it is not necessary to presuppose that Richard's violent conduct will be culpable.

Moving to probabilistic generalisations, consider the following variation on the previous example. Assume that Stephen is exposed to another type of radiation, which affects the nervous system and always causes certain skin blotches but causes an irresistible urge to attack others, when the opportunity arises, in only 80 per cent of cases. There are at least two ways to understand how this generalisation reflects the underlying causal relation between the radiation and the agent's conduct. According to the subjective interpretation of probability, which is commonly considered the most suitable for legal purposes,[9] probabilistic generalisations reflect the limited state of our knowledge rather than the true nature of the world. While the generalisation about the radiation is probabilistic, it imperfectly reflects a reality that may be deterministic. If the world is indeed deterministic, Stephen belongs to one of two possible sub-groups. One possibility is that he belongs to the sub-group of people who possess an extra unknown variable, which, together with the radiation, determines that he will go berserk. The other possibility is that he belongs to the sub-group of people who do not possess the extra variable, in which case the exposure to the radiation will not cause him to go berserk.

If Stephen possesses the extra variable, supporting his conviction by inferring from the blotches on his skin that he was (80 per cent) likely to have acted violently is problematic. Similarly to deterministic generalisations, such an inference leads to a contradiction. His conduct is taken to be both free (in order to be culpable) and unfree (as, together with another unknown variable, his violent actions were determined by the radiation). To avoid the contradiction, either the evidence of the blotches has to be accepted as probative of the violent act's having been committed by Stephen, in which case he is not culpable; or it has to be deemed not probative, in which case it should be ignored.

If Stephen does not possess the extra variable, inferring from his blotches that he was (80 per cent) likely to have acted violently is mistaken and, hence, misleading. This is because, if he belongs to the sub-group of people who were not caused to act violently by the radiation, then the probability that he acted violently is not affected by the exposure to the radiation. Inferring from the skin blotches that he is more likely to have acted violently than he would have been, had he not presented these marks, is therefore mistaken. In sum, this inference is either contradictory, because

[9] For criminal law, see Alexander et al. (2009, 31); for tort law, see Perry (1995, 333-335); for health and safety regulation, see Adler (2005, 1247).

it requires inconsistent presuppositions, or it is misleading, because it is mistaken and yet is presented as informative.

Using this evidence to support Stephen's conviction is objectionable also under the objective interpretation.[10] According to this interpretation, the radiation works in a genuinely indeterministic manner and it is impossible to know *at the time of the exposure* whether Stephen will go berserk. However, if Stephen is put to trial, the important question is whether the violent action, *which is a given*, was performed by Stephen or someone else. If the genuinely indeterministic radiation ultimately caused Stephen to go berserk, then his violent conduct was determined and not under his control. In such a scenario, the subjective and objective interpretations diverge on the question of whether the radiation, together with all relevant factors, determined Stephen's violent conduct, or whether there was room for chance. However, under both interpretations, Stephen's violent conduct was caused by a factor not under his control, and hence he was unfree and cannot be held culpable for it. By contrast, if the radiation did not ultimately cause Stephen to go berserk, then inferring from the blotches on his skin that he is likelier to have behaved violently is, again, mistaken. Therefore, inferring from the blotches that he was likelier to have acted violently is either inconsistent with his being culpable, or mistaken and hence misleading.

The culpability account is able to provide a unifying justification for the hostility of criminal fact-finding toward predictive evidence. Returning to the crime-rates scenario, for an inference from crime-rates to the resident's case to be valid, it is necessary to presuppose that there is a causal generalisation that licenses this inference, such as the dangerous character of the neighbourhood, its socio-economic conditions, and so on. Such causal factors are outside the control of the individual resident. Inferring from the crime-rates that the resident was likelier to have committed a crime involving an illegal firearm is either inconsistent with their being culpable, or mistaken. As a result, if the court draws such an inference, it implicitly concedes the presupposition that the accused did not act freely. In such a case, the court would also have to concede that the individual is not culpable (and should therefore be acquitted).[11] Alternatively, if the court seeks to avoid the implications of this inference, it ought to deem it

[10] The discussion here is based on understanding the indeterminacy of the radiation as lying in the cause itself (Lewis 1986).

[11] That convicting an accused should not be based on contradictory presuppositions should not be confused with the stronger claim that *every* case of practical decision-making is subject to *all* epistemic norms, a claim I do not endorse. Nor is it assumed that holding contradictory beliefs is, in itself, morally wrong – only that it is wrong to rely on contradictory beliefs to treat someone as culpable.

irrelevant to the individual's conduct and exclude the evidence adduced to substantiate it.

The culpability account also supports common law's traditional suspicion of previous convictions and yields some criticism of recent reforms. The rules and case law governing the admissibility of previous convictions are vast and complex, and I cannot provide a comprehensive analysis of them here. However, applying the culpability account to previous convictions of child molestation may serve as an example of how such an analysis might look. Previous convictions of child molestation are admissible in both the United Kingdom and the United States.[12] While the admission of such previous convictions has been criticised on various grounds, such as being unconstitutional (Sheft 1995), unfair (McCandless 1997, 694), and even truth-supressing (Cowley and Colyer 2010), the connection to the issue of free will seems to have gone unnoticed. The culpability account would draw attention to the importance of identifying the exact generalisation involved and considering whether using it for conviction conflicts with other presuppositions made in criminal proceedings. Like any inference about human conduct, inferring from the accused's previous convictions that they are likelier to have committed the alleged similar offence(s) relies on a causal generalisation. For example, these previous convictions may be probative because they indicate that the accused suffers from a condition, such as perversion, illness, or addiction, that raises the probability of reoffending. According to the culpability account, if these previous convictions are indeed probative, it might be at the price of exposing that the accused's conduct is unfree and thus nonculpable.

One might retort that my analysis stands in contrast to a common intuitive view of criminal responsibility. While the analysis implies that the agent's conduct is either fully determined or entirely unaffected, the practices of assigning criminal responsibility often seem to assume that an agent can be *partially* causally influenced. The agent is treated as causally influenced by some factor, but only to some degree, leaving them with a less-than-maximum extent of criminal responsibility. For example, a paedophile's sentence might be mitigated by the fact that he was a victim of molestation in his childhood. According to this view, the mitigation acknowledges that his childhood experience causally influenced the way he currently acts, yet it left him sufficiently responsible for molesting children now that he is an adult.

[12] For the United Kingdom, see the Criminal Justice Act 2003, c 44, pt 11, ch 1, s 103, and for the United States, see Rule 414 of the Federal Rules of Evidence.

The difficulty with this view of criminal responsibility is that it fails to account for the conviction stage of the trial, which seeks a binary outcome: the accused is either guilty of the alleged crime or not. Finding him guilty requires that he is culpable of committing the crime, which, in turn, requires that he acted freely. Free action is thus a precondition of criminal responsibility, and, when undermined by a defence such as insanity or duress, the accused is found *not* guilty rather than *less* guilty.

One means of explaining away the intuitive force of this view of criminal responsibility is to note that, while the question of guilt is binary, the consequences of conviction are typically scalar. The punishment could include a longer or shorter period of imprisonment or a heftier or lighter fine. It is at the sentencing stage that the paedophile's childhood experience is taken into consideration. However, there could be various explanations for why this experience serves to mitigate the appropriate punishment that make no reference to a partial causal influence. To mention just a few alternatives, there would be the increased effect that punishment would have on him as a result of his experience, his vulnerability to becoming a victim again during imprisonment, or maybe even the attempt to compensate him for his bad fortune in childhood.

Whatever the justification may be, it need not rely on a causal generalisation, according to which his childhood experience causally influenced him to commit the alleged offence. Moreover, if such a causal generalisation is used at the sentencing stage, it becomes difficult to explain why the prosecution should not be allowed to admit the very same evidence at the conviction stage to support its allegation that the accused has committed the offence. The challenge here is not only to identify a solid objection to the use of such evidence in criminal trials (which is more difficult than it might seem), but also to explain why the same objection is not equally applicable at the sentencing stage. While exploring the justification for such mitigation lies outside the scope of this paper, it suffices to note that taking into account the paedophile's childhood experiences at the sentencing stage need not be based on his being less responsible for molesting the children he did. Therefore, my analysis does not stand in contrast to current sentencing practices. On the contrary, the "partial causal influences" view stands in contrast to our binary practices of *conviction*. Proponents of such a view would thus need to explain how freedom and criminal responsibility work, in their understanding.

5. Conclusion

Accepting that many free actions are necessarily unpredictable might be viewed as implausible and counterintuitive, even by libertarians such as Lockie. Yet, it seems that Lockie's self-determinism cannot be settled with the predictability of free actions. While I tend to think that, if free will exists, it is necessarily unpredictable, I did not pursue this claim here. Rather, I suggested that a theory of free will in which free actions are necessarily unpredictable is able to provide the sought-after justification for excluding predictive evidence. So even if Lockie's theory implies that self-determined actions are necessarily unpredictable, this might not be a bad thing after all ...

REFERENCES

Adler, M. 2005. Against 'individual risk': A sympathetic critique of risk assessment. *University of Pennsylvania Law Review* 153: 1121-1150.

Alexander, L., K. Ferzan, and S. Morse. 2009. *Crime and Culpability: A Theory of Criminal Law*. Cambridge: Cambridge University Press.

Allen, R., and M. Pardo. 2008. Juridical proof and the best explanation. *Law and Philosophy* 27: 223-268.

Allen, R., and M. Pardo. 2007. The problematic value of mathematical models of evidence. *Journal of Legal Studies* 36: 107-140.

Arntzenius, F. 1992. The common cause principle. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1992: 227-237.

Bernáth, L., and J. Tőzsér. 2019. Rolling back the rollback argument. Unpublished manuscript (on file with author).

Cohen, L. 1977. *The Probable and the Provable*. Oxford: Clarendon Press.

Cowley, M., and J. Colyer. 2010. Asymmetries in prior conviction reasoning: Truth suppression effects in child protection contexts. *Psychology, Crime and Law* 16: 211-231.

Dant, D. 1988. Gambling on the truth: The use of purely statistical evidence as a basis for civil liability. *Columbia Journal of Law and Social Problems* 22: 31-70.

Enoch, D., L. Spectre, and T. Fisher. 2012. Statistical evidence, sensitivity, and the legal value of knowledge. *Philosophy and Public Affairs* 40: 197-224.

Goodman, N. 1983. *Fact, Fiction, and Forecast* (4th ed). Cambridge, MA: Harvard University Press.

Hájek, A. 2009. Fifteen arguments against hypothetical frequentism. *Erkenntnis* 70: 211-235.

Hart, H. 2008. *Punishment and Responsibility* (2nd ed). Oxford: Oxford University Press.

Koehler, J. 2002. When do courts think base rate statistics are relevant? *Jurimetrics Journal* 42: 373-402.

Lewis, D. 1986. A Subjectivist's Guide to Objective Chance. In his *Philosophical Papers* (vol. 2). Oxford: Oxford University Press.

Lockie, R. 2018. *Free Will and Epistemology: A Defence of The Transcendental Argument for Freedom*. London: Bloomsbury.

McCandless, J. 1997. Prior bad acts and two bad rules: The fundamental unfairness of federal rules of evidence 413 and 414. *William & Mary Bill of Rights Journal* 5: 689-715.

Nesson, C. 1985. The evidence or the event? On judicial proof and the acceptability of verdicts. *Harvard Law Review* 98: 1357-1392.

O'Connor, T. 2000. *Persons and Causes: The Metaphysics of Free Will*. Oxford: Oxford University Press.

O'Connor, T. 2009. Agent-Causal Power. In *Dispositions and Causes*, ed. T. Handfield, 189-214. Oxford: Oxford University Press.

Perry, S. 1995. Risk, Harm, and Responsibility. In *Philosophical Foundations of Tort Law*, ed. D. Owen, 321-346. Oxford: Clarendon Press.

Pundik, A. 2006. Epistemology and the law of evidence: Four doubts about Alex Stein's foundations of evidence Law. *Civil Justice Quarterly* 25: 504-28.

Pundik, A. 2008a. What is wrong with statistical evidence? The attempts to establish an epistemic deficiency. *Civil Justice Quarterly* 27: 461-493.

Pundik, A. 2008b. Statistical evidence and individual litigants: A reconsideration of Wasserman's argument from autonomy. *International Journal of Evidence and Proof* 12: 303-324.

Pundik, A. 2011. The epistemology of statistical evidence. *International Journal of Evidence and Proof* 15: 117-143.

Pundik A. 2017. Freedom and generalisation. *Oxford Journal of Legal Studies* 37: 189-121

Redmayne, M. 2008. Exploring the proof paradoxes. *Legal Theory* 14: 281-309.

Reichenbach, H. 1999. *The Direction of Time*. Mineola, NY: Dover Publications.

Roberts, P., and A. Zuckerman. 2010. *Criminal Evidence* (2nd ed). Oxford: Oxford University Press.

Schauer, F. 2003. *Profiles, Probabilities and Stereotypes*. Cambridge: Harvard University Press.

Schoeman, F. 1987. Statistical vs. direct evidence. *Noûs* 21: 179-198.

Sela, G. 2017. Torts as Self-Defense. DPhil Thesis. University of Oxford.
Shaviro, D. 1989. Statistical-probability evidence and the appearance of justice. *Harvard Law Review* 103: 530-554.
Sheft, M. 1995. Federal rule of evidence 413: A dangerous new frontier. *American Criminal Law Review* 33: 57-87.
Stein, A. 2005. *Foundations of Evidence Law*. Oxford: Oxford University Press.
Strawson, P. 1962. Freedom and resentment. *Proceedings of the British Academy* 48: 1-25.
Tadros, V. 2005. *Criminal Responsbility*. Oxford: Oxford University Press.
Thomson, J. 1986. Liability and individualized evidence. *Law and Contemporary Problems* 49: 199-219.
Tribe, L. 1971. Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review* 84: 1329-1393.
Uviller, R. 1982. Evidence of character to prove conduct: Illusion, illogic, and injustice in the courtroom. *Pennsylvania Law Review* 130: 845-891.
van Inwagen, P. 2000. Free will remains a mystery: The eighth philosophical perspectives lecture. *Philosophical Perspectives* 14: 1-19.
Vicens, L. 2016. Objective probabilities of free choice. *Res Philosophica* 93: 125-315.
Vigen, T. 2015. Spurious Correlations. Online at: http://www.tylervigen.com/spurious-correlations. Accessed 12/16/2019.
Wasserman, D. 1992. The morality of statistical proof and the risk of mistaken liability. *Cardozo Law Review* 13: 935-76.
Williamson, J. 2005. *Bayesian Nets and Causality*. Oxford: Oxford University Press.
Zuckerman, A. 1986. Law, fact or justice? *Boston University Law Review* 66: 487-508.

ABSTRACTS (IN CROATIAN)

BITI SPOSOBAN ILI NE BITI SPOSOBAN? TO JE PITANJE. PROBLEM ZA TRANSCENDENTALNI ARGUMENT ZA SLOBODU VOLJE

NADINE ELZEIN
The University of Sydney

TUOMAS K. PERNU
University of Helsinki i King's College London

SAŽETAK

Prema jednom tipu argumenta za libertarijansku slobodu volje ako slobodno djelovanje zahtijeva dostupnost alternativnih mogućnosti, i determinizam je istinit, tada nismo opravdani tvrditi da ne postoji slobodna volja. Preciznije: ako je djelatnik A opravdan kada tvrdi propoziciju P (npr. "ne postoji slobodna volja"), onda A mora biti u stanju tvrditi ne-P. Stoga ako A nije sposoban tvrditi ne-P, zbog determinizma, onda A nije opravdan tvrditi P. Dok se takvi argumenti često pozivaju na principe koji su mnogima privlačni, poput principa da 'treba' implicira 'može', također obvezuju na principe koji se čine puno manje uvjerljivima, npr. princip da 'treba' implicira 'sposobnost ne učiniti' ili princip da imati obvezu implicira odgovornost. Argumentiramo da su ovi principi sporni te da će biti teško osmisliti valjani transcendentalni argument bez njih.

Ključne riječi: determinizam; epistemički deontologizam; sloboda volje; libertarijanizam; normativnost; 'treba' implicira 'sposobnost ne učiniti'; 'treba' implicira 'može'; PAP; praktični deontologizam; razlozi; odgovornost; transcendentalni argumenti

DETERMINIZAM I MOĆ SUĐENJA. KRITIKA INDIREKTNOG EPISTEMIČKOG TRANSCENDENTALNOG ARGUMENTA ZA SLOBODU

LUCA ZANETTI

SAŽETAK

U nedavno objavljenoj knjizi pod naslovom *Free will and epistemology*, Robert Lockie argumentira da je vjerovanje u determinizam samopobijajuće. Lockiejev argument ovisi o tvrdnji da smo obvezani vrednovati opravdanost naših vjerovanja oslanjajući se na internalističko deontološku koncepciju opravdanja. Međutim, pobornik determinizma negira postojanje slobodne volje koja je potrebna kako bismo formirali opravdana vjerovanja prema takvoj deontološkoj koncepciji opravdanja. Prema tome, u svjetlu onoga što pobornik determinizma prihvaća, samo vjerovanje u determinizam ne može biti opravdano. Na temelju toga Lockie argumentira da smo obvezani djelovati i vjerovati pod pretpostavkom da smo slobodni. U ovom radu, razmatram i odbacujem Lockiejev transcendentalni argument za slobodu. Njegov argument pretpostavlja kako je pri donošenju suda da je determinizam istinit, pobornik determinizma obvezan smatrati da postoje epistemičke dužnosti — npr. dužnost vjerovati da je determinizam istinit ili dužnost ciljati da vjerujemo istine o determinizmu. Argumentiram da se ova pretpostavka temelji na pogrešnoj koncepciji međuodnosa moći suđenja i obveza.

Ključne riječi: epistemička deontologija; sloboda volje; transcendentalni argumenti; moć suđenja

JE LI SKEPTICIZAM U POGLEDU SLOBODE VOLJE SAMOPOBIJAJUĆI?

SIMON-PIERRE CHEVARIE-COSSETTE
King's College London

SAŽETAK

Skeptici u pogledu slobode volje negiraju postojanje slobodne volje, to jest moć upravljanja ili kontrolu koja je potrebna za moralnu odgovornost. Epikurejci tvrde da je to poricanje u nekom smislu samopobijajuće. Kako bismo blagonaklono interpretirali epikurejsku tvrdnju, prvo moramo shvatiti da propozicijski stavovi poput vjerovanja

mogu biti samopobijajući, a ne same propozicije koje čine sadržaj tih stavova. Dakle, vjerovanje u skepticizam u pogledu slobodne volje može biti samopobijajući. Optužba postaje uvjerljivija jer, kao što je Epikur pronicljivo prepoznao, postoji jaka veza između ponašanja i vjerovanja te, stoga, između sadržaja skepticizma u pogledu slobode volje (budući da se radi o ponašanju) i stava koji uključuje vjerovanje u njega. Drugo, moramo shvatiti da stav može biti samopobijajući s obzirom na određene razloge. To znači da može biti samopobijajuće biti skeptik u pogledu slobode volje s obzirom na određene razloge, poput pretpostavljene činjenice da nam nedostaje slobodnog prostora za djelovanje ili izvornosti. Ta optužba je mnogo zanimljivija zbog epistemičke važnosti slobodnog prostora za djelovanje ili izvornosti. U konačnici, epikurejska optužba samopobijanja nije uspješna. Ipak, ona skepticima donosi važne lekcije. Najvažnija od njih je da skeptici u pogledu slobode volje trebaju ili prihvatiti postojanje slobodnog prostora za djelovanje ili odbaciti princip da "treba" implicira "može".

Ključne riječi: skepticizam u pogledu slobode volje; samopobijanje; slobodni prostor; izvornost; Epikurej; "treba" implicira "može"; odgovornost; razlozi

KAKO ZNAMO DA SMO SLOBODNI?

TIMOTHY O'CONNOR
Indiana University

SAŽETAK

Prirodno smo skloni vjerovati o sebi i drugima da smo slobodni: da ono što radimo često, i u velikoj mjeri, "ovisi o nama" kroz upražnjavanje moći izbora da učinimo ili se suzdržimo od poduzimanja jedne ili više alternativa kojih smo svjesni. U ovom članku istražujem izvor i epistemičko opravdanje našeg "uvjerenja u slobodu". Predlažem teoriju koja se (za razliku od većine) ne oslanja previše na naše osobno iskustvo izbora i djelovanja te umjesto toga uzima vjerovanje u slobodu kao a priori opravdano. Nakon toga, razmotrit ću moguće odgovore dostupne kompatibilistima na tvrdnju nekih kompatibilista da "privilegirani" epistemički status uvjerenja o slobodi (koji moja teorija prihvaća) podržava minimalistički, a samim time i kompatibilistički pogled na prirodu same slobode.

Ključne riječi: sloboda volje; iskustvo; nekompatibilizam; a priori opravdanje; svjesnost; revizionizam

POJMOVNA NEMOGUĆNOST TEORIJE POGREŠKE O SLOBODI VOLJE

ANDREW J. LATHAM
The University of Sydney

SAŽETAK

Ovaj rad zalaže se za gledište o slobodi volje koje ću nazvati pojmovna nemogućnost teorije pogreške o slobodi volje - teza o pojmovnoj nemogućnosti. Argumentirat ću da s obzirom na pojam slobodne volje kakav stvarno koristimo, nije moguće da su naši sudovi o slobodnoj volji - prosudbe o tome je li neka radnja slobodna ili ne - sustavno neistiniti. Budući da za mnoge od naših postupaka smatramo da su slobodni, iz teze o pojmovnoj nemogućnosti slijedi da su mnogi naši postupci stvarno slobodni. Iz toga proizlazi da je teorija pogreške o slobodnoj volji - gledište da nijedna prosudba koja ima formu "radnja A je slobodno izvršena" - netočna. Prikazat ću da ozbiljno uzimanje u obzir teze o pojmovnoj nemogućnosti pomaže učiniti smislenim neke od naizgled nekonzistentnih rezultata iz eksperimentalne filozofije o determinizmu i pojmu slobode volje. Nadalje, navest ću neke razloge zašto bismo trebali očekivati da ćemo pronaći slične rezultate za svaki drugi faktor za koji smatramo da je važan za slobodu volje.

Ključne riječi: sloboda volje; teorija pogreške; pojmovna nemogućnost; uvjetni pojam; eksperimentalna filozofija

MOŽEMO LI PREDVIDJETI SAMOUVJETOVANE RADNJE?

AMIT PUNDIK
Tel Aviv University

SAŽETAK

Ovaj rad ispituje Lockiejevu teoriju o libertarijanskom samoodređivanju u svjetlu pitanja koje se odnosi na mogućnost predviđanja: "Možemo li znati (ili opravdano vjerovati) kako će slobodan djelatnik djelovati ili vjerojatno djelovati?" Argumentiram da, kad se Lockiejeva teorija uzme do logički krajnjih granica, slobodne radnje se ne mogu predvidjeti u bilo kojoj mjeri točnosti, čak i ako imaju vjerojatnosti, one se ne mogu znati. No, predlažem da je ta implikacija njegove teorije zapravo korisna jer može objasniti i opravdati važnu karakteristiku praksi koje koristimo za utvrđivanje je li netko kriv za određeni postupak: naš negativan stav prema upotrebi prediktivne dokazne građe.

Ključne riječi: sloboda volje; uzročnost; objektivna vjerojatnost; determinizam; krivična odgovornost; Dennett; predviđanje; Lockie

Prijevod: Marko Jurjako i Jelena Kopajtić

INSTRUCTIONS FOR AUTHORS

Publication ethics

EuJAP subscribes to the publication principles and ethical guidelines of the Committee on Publication Ethics (COPE).

Submission

Submitted manuscripts ought to be:

- be unpublished, either completely or in their essential content, in English or other languages, and not under consideration for publication elsewhere;
- be approved by all co-Authors;
- contain citations and references to avoid plagiarism, self-plagiarism, and illegitimate duplication of texts, figures, etc. Moreover, Authors should obtain permission to use any third party images, figures and the like from the respective copyright holders. The pre-reviewing process includes screening for plagiarism and self-plagiarism by means of internet browsing;
- be sent exclusively electronically to the Editors (eujap@ffri.hr) (or to the Guest editors in the case of a special issue) in a Word compatible format;
- be prepared for blind refereeing: authors' names and their institutional affiliations should not appear on the manuscript. Moreover, "identifiers" in MS Word Properties should be removed;
- be accompanied by a separate file containing the title of the manuscript, a short abstract (not exceeding 250 words), keywords, academic affiliation and full address for correspondence including e-mail address, and, if needed, a disclosure of the Authors' potential conflict of interest that might affect the conclusions, interpretation, and evaluation of the relevant work under consideration;
- in American or British English
- no longer than 8000 words, including footnotes and references

The Editors reserve the right to reject submissions that do not satisfy any of the previous conditions.

If, due to the authors' failure to inform the Editors, already published material will appear in EuJAP, the Editors will report the authors'

unethical behaviour in the next issue and remove the publication from EuJAP web site and the repository HRČAK.

In any case, the Editors and the publisher will not be held legally responsible should there be any claims for compensation following from copyright infringements by the authors.

If the manuscript does not match the scope and aims of EuJAP, the Editors reserve the right to reject the manuscript without sending it out to external reviewers.

Style

Accepted manuscripts should:

- follow the guidelines of the most recent Chicago Manual of Style
- contain footnotes and no endnotes
- contain references in accordance with the author-date Chicago style, here illustrated for the main common types of publications (T = in text citation, R = reference list entry)

Book
T: (Nozick 1981, 203)
R: Nozick, R. 1981. *Philosophical Explanations.* Cambridge: Harvard University Press.

Chapter or other part of a book
T: (Fumerton 2006, 77-9)
R: Fumerton, R. 2006. The Epistemic Role of Testimony: Internalist and Externalist Perspectives. In *The Epistemology of Testimony*, ed. J. Lackey and E. Sosa, 77-92. Oxford: Oxford University Press.

Edited collections
T: (Lackey and Sosa 2006)
R: Lackey, J. and E. Sosa, eds. 2006. *The Epistemology of Testimony*. Oxford: Oxford University Press.

Article in a print journal
T: (Broome 1999, 414-9)
R: Broome, J. 1999. Normative requirements. *Ratio* 12: 398-419.

Electronic books or journals
T: (Skorupski 2010)

R: Skorupski, J. 2010. Sentimentalism: Its Scope and Limits. *Ethical Theory and Moral Practice* 13: 125-136. http://www.jstor.org/stable/40602550

Website content
T: (Brandon 2008)
R: Brandon, R. 2008. Natural Selection. *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta. Accessed September 26, 2013. http://plato.stanford.edu/archives/fall2010/entries/natural-selection

Forthcoming
For all types of publications followed should be the above guideline style with exception of placing 'forthcoming' instead of date of publication. For example, in case of a book:
T: (Recanati forthcoming)
R: Recanati, F. forthcoming. *Mental Files*. Oxford: Oxford University Press.

Unpublished material
T: (Gödel 1951)
R: Gödel, K. 1951. *Some basic theorems on the foundations of mathematics and their philosophical implications*. Unpublished manuscript, last modified August 3, 1951.

Final proofreading

Authors are responsible for correcting proofs.

Copyrights

The journal allows the author(s) to hold the copyright without restrictions. In the reprints, the original publication of the text in EuJAP must be acknowledged by mentioning the name of the journal, the year of the publication, the volume and the issue numbers and the article pages.

EuJAP subscribes to Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). Users can freely copy and redistribute the material in any medium or format, remix, transform, and build upon the material for any purpose. Users must give appropriate credit, provide a link to the license, and indicate if changes were made. Users may do so in any reasonable manner, but not in any way that suggests the licensor endorses them or their use. Nonetheless, users must distribute their contributions under the same license as the original.

Archiving rights

The papers published in EuJAP can be deposited and self-archived in the institutional and thematic repositories providing the link to the journal's web pages and HRČAK.

Subscriptions

A subscription comprises two issues. All prices include postage.

Annual subscription:

International:

individuals € 10

institutions € 15

Croatia:

individuals 40 kn

institutions 70 kn

Bank: Zagrebačka banka d.d. Zagreb

SWIFT: ZABAHR 2X
IBAN: HR9123600001101536455

Only for subscribers from Croatia,

please add: “poziv na broj”: 0015-03368491

European Journal of Analytic Philosophy is published twice per year.

The articles published in the European Journal of Analytic Philosophy are indexed and abstracted in The Philosopher's Index, European Reference Index for the Humanities (ERIH PLUS), Directory of Open Access Journals (DOAJ), PhilPapers, and Portal of Scientific Journals of Croatia (HRČAK).