Resisting the epistemic argument for compatibilism

Resisting the Epistemic Argument for Compatibilism

Patrick Todd and Brian Rabern

Forthcoming, *Philosophical Studies*

Philosophers are keenly aware that there are at least several (much-discussed) arguments that free will and moral responsibility are incompatible with determinism. But whereas the standard arguments for incompatibilism are well-known, arguments of the opposite sort – positive arguments *for* compatibilism – are correspondingly harder to come by. In general, it seems that the main way that philosophers defend compatibilism is simply by trying to show that the arguments *for* incompatibilism fail. And perhaps this makes sense. After all, one might think that there is, in general, some sort of presumption of compatibility; if someone asks for a reason to think that the fact that there is water on Earth is compatible with the fact that there is also water on Mars, the natural reaction is presumably to say: well, why *shouldn't* these things be compatible? Similarly, to pick two philosophical positions at random, if one were asked for some sort of argument that consequentialism in ethics is compatible with reliabilism in epistemology, presumably one would find it difficult to know what to say. Proving that two things are compatible sometimes can be nothing more than rebutting any argument that they aren't.

But even if there *is* this kind of presumption in favour of compatibilism, at least some philosophers have indeed given positive arguments for compatibilism. In this paper, we focus on what certainly seems to be the most prominent such argument: what we propose to call the *epistemic argument* for compatibilism. Epistemic arguments of the relevant kind proceed from epistemic premises – premises about our knowledge or evidence – to substantial metaphysical conclusions. As we will see below, our primary focus will be on one such argument as articulated by David Lewis. But the basic intuitive machinery underlying Lewis' argument can plausibly be found, in various guises, in almost all of contemporary compatibilist thinking. For instance, an epistemic argument is plausibly at work in Peter Strawson's famous insistence that "the facts as we know them" are a sufficient basis for our responsibility practices – and that since the facts as we know them do not rule out determinism, it cannot be that our responsibility practices require the falsity of determinism.[1] Though our primary focus will be on Lewis' argument for *ability to do*

[1] Strawson (1962: 208); cf. Coates' (2017: 820) interpretation of Strawson: "First, because all but the most hardened sceptics will admit that we already know our general engagement with others as friends, lovers, colleagues, and even parties to chance encounters does not stand in need of further justification, the legitimacy of these relationships cannot depend on the falsity of causal determinism, since we do not currently know causal determinism to be false."

otherwise compatibilism (what we shall here mean by "free will"), the key responses we develop can, we believe, be applied to epistemic arguments of other sorts.

Call the following two premises "the data" underlying the argument:

(1) We know we are free.

(2) For all we know everything is predetermined.

Of course, the data stands in need of interpretation – for instance, to whom does "we" refer? As a first approximation, however, the epistemic argument in question somehow moves from the data to a compatibilist conclusion. Now, there will be, of course, responses to this style of argument that outright deny what its proponents consider "the data". For instance, skeptics about freedom will simply deny (1). And certain libertarians will outright deny (2). In this paper, however, we set aside simplistic data-denying responses. Instead, we aim to show that compatibilism doesn't follow from a natural interpretation of the data in question.

We proceed as follows. We begin by articulating Lewis' epistemic argument, highlighting its crucial features, and avoiding certain key interpretive pitfalls, while setting aside what we take to be weak responses to the argument. We then turn to our preferred assessment of the argument. In doing so, we will draw out a connection between the epistemic argument and parallel issues that have arisen in the physicalism/dualism debates. If a dualist concedes (as it seems they should) that for all they know an oracle might tell them that the world is merely physical, then, since they know they are conscious, this concession would seem to force them into abandoning their commitment to dualism. But we will show that, properly understood, this concession is no threat to the dualist. And, in a parallel fashion, we will argue that an incompatibilist needn't be threatened by the concession underwriting (2), viz., that science might tell us that the world is completely deterministic. We contend that this makes sense of the *flip-flopping* strategy famously exemplified by Peter van Inwagen (1983). Van Inwagen maintains that incompatibilism is true (and that we are free) – but that if we somehow got decisive empirical reason to think that determinism is true, he would give up his (*a priori*) belief in incompatibilism, rather than his belief in freedom. Fischer (2016) has argued that this kind of flip-flopping is unstable.[2] We aim show that it isn't, and that it affords a principled way of resisting the epistemic argument. The resulting libertarian position, however, does encounter the worry that we can (in some sense) rule out determinism "from the armchair" (Fischer 2007: 46 – 7). We conclude by

[2] As far as we can see, the first place Fischer makes the "flipflopping" charge is together with Ravizza in Fischer and Ravizza 1998: 253 – 4; cf. also Fischer 2007: 46 – 7. Fischer 2016, however, is exclusively focussed on developing and expanding this complaint.

articulating two different libertarian responses to this complaint. The result is that the dialectic concerning the epistemic argument arguably needs to shift towards the evaluation of these libertarian strategies.

1. Lewis' epistemic argument

Lewis' only (non-posthumously) published statement of his epistemic argument for compatibilism is admirably brief, and it goes like this:

> The best argument for compatibilism is that we know better that we are sometimes free than that we ever escape predetermination; wherefore it may be for all we know that we are free but predetermined. (Lewis 1993: 155)

Two things to initially note about this argument. First, not only is the argument brief – it is also enthymematic. The conclusion is supposed to be compatibilism, but all that is said to follow from the explicit premise is *epistemic* compatibilism, i.e., *for all we know* we are free but predetermined. Thus, Lewis must think that *metaphysical* compatibilism is an immediate consequence of *epistemic* compatibilism. Second, Lewis' premise is of the form

> x knows better that p than that q

and according to some, this construction is ungrammatical (Stanley 2004: 123-30). But while the question of the gradeability of the verb "know" is interesting, we'll side-step it here, because it seems clear what Lewis meant, even if his syntax is controversial.

In his posthumously published paper "Nihil Obstat", Lewis gives essentially the same argument while making the implicit premise more explicit and without using the "know better" locution. He calls it a *simple proof* of compatibilism (Lewis 2020: 241):

> i. It's a Moorean fact that we often have a choice what to do.
> ii. But whether determinism holds is an unsettled question.
> iii. So having a free choice is epistemically compatible with determinism.
> iv. So, it's compatible *simpliciter*.

If something is a "Moorean fact" then it is epistemically more secure than "the premises of any philosophical argument to the contrary."[3] And if whether something holds is "unsettled", then for all we know it holds. This provides good evidence that what Lewis meant (or at least what he took to follow from) his "know better" construction is essentially the conjunction of (i) and (ii).[4] The inference from (iii) to (iv) also provides good evidence that Lewis indeed took (metaphysical) compatibilism to follow from epistemic compatibilism. Thus, we propose to regiment Lewis' epistemic argument as follows:

(1) We know that we are free.
(2) For all we know everything is predetermined.
(3) If we know that we are free but for all we know everything is predetermined, then for all we know we are free but everything is predetermined.
(4) If for all we know we are free but everything is predetermined, then being free is compatible with being predetermined.
(5) So, being free is compatible with being predetermined.

The argument looks valid. Before assessing it in detail let's take it at face value and walk through a simple model of the premises in Lewisian terms.

Lewis advocated a modal conception of knowledge and belief (see Lewis 1986: 27-39; cf. Hintikka 1962). The basic idea is that to have knowledge is to locate the actual world in the space of all possible worlds. Given all the ways the world could be, some of those ways are compatible with our evidence, while other are incompatible. So, if we know that *p*, then our evidence rules out all the ~*p*-worlds as candidates for actuality. That is, all the worlds left uneliminated by our evidence are *p*-worlds. These remaining epistemic possibilities are a subset of all the possibilities.

Now determinism is a contingent thesis. It holds in some worlds but not in others. Premise (2) says that for all we know we are predetermined. So, the worlds left uneliminated by our evidence must include some deterministic worlds. Premise (1) says we know we are free, so in all the worlds left uneliminated by our evidence we are free. Given this setup the conclusion

[3] Lewis (1996: 549); Lewis also approvingly cites Armstrong saying that a Moorean fact is "one of the many facts that even philosophers should not deny, whatever philosophical analysis they give of such facts" (1999: 20). See Nolan (2015) for discussion of the role of Moorean facts in Lewis' methodology.
[4] In fact, Lewis (1996: 562-3) says that "better knowledge" is *more stable* knowledge – knowledge that rests more on the elimination of possibilities rather than the ignoring of them. To truly say "We know better that we are sometimes free than that we ever escape predetermination," we must be attending to some uneliminated possibilities in which we never escape predetermination – and in this case we could truly say both "We know that we are sometimes free" and "We might never escape predetermination". See Lewis (1996: footnote 19).

follows almost immediately. Premise (3) says it then follows that we might be free but predetermined, which is to say that there is some possible world where we are free but predetermined left uneliminated by our evidence. Premise (4) says if that is so, then being free is compatible with being predetermined, which is to say there is a possible world where we are both free and predetermined. So, compatibilism. In terms of Lewis' preferred possible-worlds account of knowledge and modality, the argument seems fairly straightforward.[5]

However, one initially tempting reply that we want to set aside is an appeal to *epistemic contextualism*, à la Lewis (1996).[6] One could suggest that perhaps (1) is true in some contexts and (2) is true in others, but in contexts in which (2) is true, (1) isn't. With contextual variability in the picture, the inference from "the data" to the compatibilist conclusion is rendered invalid; in particular, the consequent of premise (3) wouldn't follow from (1) and (2). But notice that the natural way to implement this response yields that we – us philosophers thinking about determinism, who concede that we can't rule it out – don't know that we are free! Thus, one might worry that this reply concedes too much to the freedom skeptic. For this reason, we won't pursue this line of response here. In particular, we aren't denying the context-sensitivity of "know"; instead, we are granting the proponent of the epistemic argument a robust understanding of premise (1), i.e., we are granting that we (us philosophers) truthfully say "We know we are free" even in contexts in which the epistemic possibility of determinism is under discussion (or is "relevant" or "attended to"; see footnote 3 above). In general, we are granting that our freedom is epistemically secure, it's undeniable, it's Moorean, and so on. We thus wish to grant that the consequent of (3) follows from (1) and (2).

But now let us consider premise (4).[7] In general, the form of this premise is as follows: If for all we know *p*, then it is genuinely possible that *p*. Now, there would appear to be false

---

[5] Given a standard Kripke semantics for a multi-modal logic with epistemic ($\blacklozenge$, $\blacksquare$) and alethic ($\Diamond$, $\Box$) modals, the conclusion would be entailed by the premises (assuming at least D for $\blacksquare$):

$\blacksquare f, \blacklozenge d, (\blacksquare f \wedge \blacklozenge d) \rightarrow \blacklozenge (f \wedge d), \blacklozenge (f \wedge d) \rightarrow \Diamond (f \wedge d) \vDash \Diamond (f \wedge d)$.

[6] Note that the contextualism alluded to here is different from Hawthorne's (2001) contextualism about "freedom". That account is importantly not about attributions of *knowledge* of freedom, it is about *freedom* claims themselves. Hawthorne would presumably also deny that we truthfully say "We know we are free", since our utterances (in philosophical contexts where determinism is salient) of "We are free" are themselves false, even if true in other contexts.

[7] Chevarie-Cossette (2021) insists that Lewis' argument is unsound because – as we have rendered it – premise (4) is false. Instead, he argues that a more subtle argument for a weaker conclusion succeeds. This argument replaces (4) with the following: If for all we know we are free but everything is predetermined, then *for all we know* being free is compatible with being predetermined. And, thus, the conclusion of this more subtle argument is that *for all we know* compatibilism holds. As he says, "Lewis's argument supports the unknowability of incompatibilism, given that we know we are responsible, not the truth of compatibilism" (Chevarie-Cossette 2021: 205). This argument for a weaker conclusion – that incompatibilism (and thus also libertarianism) can't be *known* – raises some different issues from those with which we are mainly concerned here. But as will become apparent below, given the way we interpret the

instances of this schema. Lewis insists, however, that these sorts of (in his words) "impossible epistemic possibilities" fall into a few special classes, but that the case at issue doesn't fall into one of those classes (cf. Beebee et. al. 2020 for discussion). In short, according to Lewis, counterexamples to the given schema are either cases of (a) the necessary *a posteriori*, or cases of (b) *mathematical/modal ignorance*.[8] Unfortunately, Lewis doesn't indicate why exactly he takes it that these cases are irrelevant – so we will have to fill in some details.

Cases involving the necessary *a posteriori* are fairly clear. Consider the canonical example. Water is composed of H2O. But there was a time when for all we knew water didn't contain hydrogen. It doesn't follow, however, that it was metaphysically possible for there to be water without hydrogen. As Kripke (1980) has taught us, it can be epistemically possible that water doesn't contain hydrogen (or that cats are robots, or that Hesperus isn't Phosphorus) without it being genuinely (metaphysically) possible that water doesn't contain hydrogen (or that cats are robots, or that Hesperus isn't Phosphorus). Now, if this precedent is irrelevant, then it has to be that "free will" (or whatever concept is at issue in the epistemic argument in question) is relevantly dissimilar to "water" or "cats". Whether the existence of water is incompatible with the absence of H2O presumably depends on the underlying nature of water, which must be discovered empirically. But whether freedom is incompatible with determinism does not seem to depend on some *empirical* discovery; that question would instead appear to be *a priori*. Thus, Lewis insists that case (a) examples are irrelevant. If incompatibilism is true at all, then its truth is *a priori*.

Look at it this way. The reason why the epistemic possibility that water doesn't contain hydrogen does not entail the *real* possibility that water doesn't contain hydrogen has something to do with the special status of a term like "water" – in particular, it has to do with the role the *external environment* plays in fixing its meaning. What such an expression picks out in counterfactual worlds depends on how things in the actual environment turn out. Expressions that have this sort of feature are sometimes called "twin-earthable", or semantically unstable.[9]

---

best version of Lewis' argument (i.e. in terms of conceivability), the subtle argument (so interpreted) would also arrive at Lewis' stronger conclusion.

[8] Lewis (2020) actually lists three "alleged precedents": (i) mathematical or logical ignorance, (ii) the geography of the pluriverse, and (iii) necessity *a posteriori*. Since Lewis often talks about mathematical and modal ignorance under the same heading, we will simplify by grouping (i) and (ii) here (see Lewis 1986: 108-15).

[9] Chalmers (2006) makes this sort of distinction within his preferred two-dimensional framework: a non-twin-earthable (or neutral) expression is one whose extension in counterfactual worlds does not depend on how the actual world turns out. Bealer (1996) makes a similar distinction but without the two-dimensional apparatus in terms of what he calls "semantic stability". An expression is semantically *stable* just in case, necessarily, in any language group in an epistemic situation qualitatively identical to ours, the expression would have the same meaning; and an expression is unstable otherwise (Bealer, 1996: 134). According to Bealer ,"water" is unstable, while "consciousness" and "freedom" are stable.

Twin-earthable expressions would (standardly) include *natural kind terms* such as "water", "tiger", and "gold", and *proper names,* such as "Hesperus" and "Gödel". But "freedom" (or, again, whatever term is at issue in the epistemic argument in question), along with "consciousness", "knowledge", "goodness", etc. would seem to be relevantly dissimilar to "water" – that is, dissimilar in their twin-earthability.[10]

It is less clear, however, why Lewis took case (b) examples to be similarly irrelevant. The proposition that we are free but predetermined is certainly neither a mathematical nor a modal proposition – but why exactly does this difference make a difference? One natural thought appeals to a difference in modal profile. Consider the *twin prime conjecture*, which states that there are infinitely many primes m such that $m + 2$ is also prime. Call this proposition T. The question over T is an open problem in number theory – for all we know it might be that T and also for all we know it might be that ~T. But the answer here, whether it is T or ~T, is (let's assume) a necessary truth. So, it's a necessary truth that is epistemically open. But this case is, arguably, disanalogous to the case relevant for compatibilism. The twin primes case is one in which either T or ~T is metaphysically impossible, but we don't know which it is. But notice that the proposition relevant for the epistemic argument – that is, the proposition that we are free but predetermined – isn't like this. Here instead we have a proposition that if true is merely contingently true. Whether determinism holds is contingent, and it isn't even necessary that we exist, let alone necessary that we are free. So, this proposition doesn't fit the paradigm of a proposition p that is epistemically open, but where one of p or $\sim p$ is metaphysically impossible. Of course, according to the incompatibilist, the relevant proposition is indeed impossible. However, to insist on this is plausibly therefore to insist that the proposition is *not* epistemically open. (More on this to come.)

We aren't entirely sure whether this difference is the difference Lewis had in mind; indeed, Lewis' views on mathematical and modal ignorance are a matter of controversy (see

[10] One could, and some in fact already have (e.g., Heller 1996), likewise argue that "freedom" is twin-earthable. The idea would be that if it turns out that the relevant human behavior is suitably indeterministic, then "freedom" picks out whatever states play the freedom-role, perhaps the *libertarian powers*. But if it turns out that such behavior is deterministic, then "freedom" picks out compatibilist-freedom, whatever that is. (See Daw and Alter (2001), and Balaguer (2010: 37ff) for a number of objections.) Latham (2019) and Deery (2021) have recently defended views along these lines in defense of compatibilism. But even if there is some motivation to accept that "freedom" is twin-earthable, this is not, in the end, a promising way for the *incompatibilist* to go. Standardly, incompatibilists don't claim that their view is supported by an empirical investigation into the nature of human action – instead they put forward *a priori* arguments, e.g., the Consequence Argument (van Inwagen 1983, Speaks 2011), or the manipulation argument (Pereboom 2001: Ch.4, Pereboom 2014: Ch. 4, Todd 2017, Todd 2019, Mele 2019).

Schwarz 2022). Ultimately, however, we will suggest that cases involving mathematical and modal ignorance are more relevant than Lewis seems to have granted.

But let's slow down. Plainly, there exists a range of difficult and highly contested issues surrounding premise (4). For instance, given a modal epistemology whereby conceivability-possibility links are simply severed, there is no issue whatsoever with accepting that a proposition can't be ruled out, while nevertheless insisting that it is impossible. In this paper, however, we will grant at least a moderate form of what has been called "modal rationalism". (For discussion, see, e.g., Yablo 1993 and Chalmers 2002). In general, if certain conditions are met, and all else is equal, a *sort* of epistemic possibility does entail possibility *tout court*. But it is precisely these conditions – and the specific sort of epistemic possibility – that will become important as we proceed.

We take the above to provide a charitable exposition of the epistemic argument. Now we turn to investigate one promising way of resisting the argument.

2. The flip-flopping dualist

To make our case, we first wish to investigate a parallel issue that has arisen in the physicalism/dualism debates. According to physicalism, our conscious states are nothing over and above the physical states of the world. According to dualism, on the other hand, our conscious states *are* something over and above the physical—that is, it is a commitment of dualism that any minimal physical duplicate of our world lacks consciousness. Notably, one could seemingly give an epistemic argument against dualism that parallels Lewis' argument against incompatibilism. Consider:

(1') We know that we are conscious.
(2') For all we know everything is physical.
(3') If we know that we are conscious but for all we know everything is physical, then for all we know we are conscious but everything is physical.
(4') If for all we know we are conscious but everything is physical, then it is possible that we are conscious but everything is physical.
(5') So, it is possible that we are conscious but everything is physical [i.e., dualism is false].

The upshot: if dualists concede that it might turn out that the world is merely physical, then it seems this argument forces them into abandoning their dualism.[11]

How should the dualist respond?[12] It looks like the dualist must either deny (1') – which is plausibly a nonstarter – or instead maintain what would seem to be the hubristic position of denying (2'). One way of thinking about the (alleged) hubris involved in denying (2') is by considering Hawthorne's thought experiment involving an oracle. Hawthorne writes:

> …suppose an oracle tells you [the dualist] tomorrow that the world is merely physical. Will you conclude that there is no pain, that your earlier self was making a mistake in ascribing pain to himself on occasion? No. You will remain convinced that you do feel pain sometimes and will reckon as pain whatever plays the pain role. (Hawthorne 2002: 26)

Hawthorne is, of course, assuming that no dualist will say, in response to his thought experiment, "Well, I am totally sure that no oracle is going to tell me that!" That is, when Hawthorne puts forward this thought experiment, he is assuming that the dualist will agree that the thought experiment is epistemically possible. We can't completely rule out the possibility that an oracle is going to tell us that the world is physical! But if we can't, then for all we know, everything may be physical – in which case, we are granting premise (2').

We want to suggest that an adequate response to this argument has been provided by David Chalmers (Chalmers 2010; see also Alter 2007). Chalmers takes dualism (here understood as conscious-but-merely-physical incompatibilism) to be an *a priori* truth – but if the oracle told him that everything is physical, he would abandon his *a priori* conviction that dualism is true. That is, rather than conclude that he simply isn't conscious, Chalmers would instead conclude that his *a priori* arguments for dualism must have gone wrong somewhere, even if he can't say where. In spite of this concession, however, Chalmers retains his *a priori* conviction in dualism. To employ some terminology from the free will debate that will become importantly shortly, it looks like Chalmers is a *flip-flopper*. As we see it, however, this is a principled and coherent stance in reply to the epistemic argument for physicalism. And we will suggest that the same holds for

---

[11] Cf. Frankish (2007) on the "anti-zombie" argument.

[12] We will assume that it is not open for the dualist to insist that (4') should be rejected on the grounds that "consciousness" is twin-earthable. That is, we will assume that if true, it is not a necessary *a posteriori* truth that consciousness is non-physical. Some physicalists such as Braddon-Mitchell (2003) insist that "consciousness" *is* unstable in the requisite way (i.e., a conditional concept), but this stance, we assume, isn't useful for the dualist (see Chalmers 2010: 158-159).

the parallel stance that the incompatibilist can take to resist the epistemic argument for compatibilism.

But let's back up. Precisely which premise of the above argument does (or should) the dualist deny? As we see it, the answer to this question is subtle. As it stands, the dualist should say that the argument *equivocates*. More particularly, the dualist should contend that there are (at least) *two* readings of the relevant argument. On one such reading, though premise (2') is plausible, there is no reason to accept premise (4'). And on the other such reading, though (4') is plausible, there is no reason to accept (2'). The two readings in question correspond to the two salient interpretations of the key phrase, "for all we know". And here the dualist should insist that, in this context, this key phrase could mean either of the following (cf. Alter 2007: 240-41):

(A) For all we know *with certainty*
(B) For all we know *by means of ideal rational reflection*

These interpretations concern different, though often conflated, sorts of epistemic modality: *certainty* versus *a priority*. To say that something holds *for all we know with certainty* is simply to say that we aren't certain that it is not the case. However, to say that something holds *for all we know by means of ideal rational reflection* is to say that given full consideration, free from certain cognitive limitations, the proposition is rationally consistent and coherent. That is, it is *conceivable* on idealised reflection. Now the point. Some propositions we can't rule out with certainty – even propositions we can't rule out with certainty after substantial rational reflection – may nevertheless be ruled out after ideal *a priori* reasoning. For example, consider again the twin primes conjecture. We can't rule out the conjecture with certainty, but if (unbeknownst to us) there is a counterexample to that conjecture, then it is not ideally conceivable, as ideal reflection would rule it out. It would be epistemically possible for all of which we are certain, but not epistemically possible for all we know by means of ideal rational reflection.

Now the key thought. In order for premise (4') to be plausible, the epistemic modality must be understood as in (B). However, in order for premise (2') to be plausible, the relevant modality must be understood as in (A). In what follows, then, we disambiguate the argument in these two key ways, and show how, understood in *either* way, the argument plausibly fails.

3. The argument disambiguated

First, consider the argument disambiguated in terms of *certainty*:

(A1') We are certain that we are conscious.

(A2') For all we know with certainty everything is physical.

(A3') If (A1') and (A2'), then for all we know with certainty we are conscious but everything is physical.

(A4') If for all we know with certainty we are conscious but everything is physical, then it is possible that we are conscious but everything is physical.

(A5') So, it is possible that we are conscious but everything is physical.

The dualist can agree that the argument is valid, and can agree with (A1') – (A3'). However, the dualist may plausibly contend that we have little reason to accept (A4'). It is indeed plausible that a *sort* of epistemic possibility – or conceivability – entails genuine possibility. But this isn't just *any* sort of "conceivability". Instead, the plausible position in the neighbourhood is that it is *ideal rational* conceivability, if anything, that entails possibility (cf. Yablo 1993 and Chalmers 2002). But the epistemic modality at issue in the argument above is not ideal conceivability. Instead, it is *certainty* (or really the dual of certainty). And there is little reason to grant a move from *for all of which we are certain p* to *it is genuinely possible that p.*

It is here that the parallel with mathematical/modal ignorance once again comes into view. Suppose we think Fermat's last theorem holds; suppose we've read about the mathematician Andrew Wiles' secret multi-year effort to prove it, and about how the mathematical community accepted the proof. But we haven't confirmed the proof for ourselves. Thus, for all we know with certainty, the "theorem" might be false. Is there now some pressure to grant that it is indeed (metaphysically) possible that it is false? Well, hardly. Further, even if Wiles seems to have found a proof of the conjecture, presumably he should still admit that the god of mathematics *might* tell him that the "proof" is flawed and that Fermat's conjecture is, in fact, false. Thus, there is some sort of epistemically possible scenario in which Wiles finds out that the conjecture is false, and in such a scenario he'd give up his *a priori* belief in the conjecture. But we all should concede that we might be wrong with regard to our *a priori* convictions. As van Inwagen says, "*a priori* convictions are as corrigible as any others." (1983: 221) In this light, accepting (A2') is just a form of epistemic humility – but this concession in no way supports the conclusion that the relevant proposition is a genuine possibility.

Thus, let us take this lesson to heart: if the argument is to be plausible, the key epistemic modality underlying (4') must be ideal rational conceivability, or *conceivability*, for short. Indeed, let us now investigate the argument under the second disambiguation noted above. In order to avoid the potential charge of equivocation, let us employ the same interpretation of the given epistemic modality throughout, to wit:

(B1') It is inconceivable that we aren't conscious.

(B2') It is conceivable that everything is physical.

(B3') If (B1') and (B2'), then it is conceivable that we are conscious but everything is physical.

(B4') If it is conceivable that we are conscious but everything is physical, then it is possible that we are conscious but everything is physical.

(B5') So, it is possible that we are conscious but everything is physical.

There is, however, an obvious problem with this argument. The first premise is plainly false. It is indeed conceivable that "we" aren't conscious, for the simple reason that it is conceivable that there should have never been conscious beings in the first place. In other words, we know *a posteriori* (e.g. via introspection) that we are conscious, not *a priori*.

How then should we interpret the argument?[13] We suggest the following strategy. It is well-known that modals such as "might" or "must" are sensitive to context and background information.[14] And even holding fixed that the modality involved is *epistemic*, there are different sorts of epistemic modalities relative to different sets of evidence or bodies of information. A claim to the effect that "such-and-such must be that case" may be true relative to some evidence yet false when bracketing that evidence. So, we should ask the proponent of the epistemic argument: epistemically necessary *given what evidence*? And here the proponent of the argument might naturally suggest:[15]

(*) given our total evidence, including our *a posteriori evidence*

This evidence, of course, includes the *a posteriori* evidence that we are conscious. So, on this approach, the key thought behind premise (1') is not the claim that we know *a priori* that we are

---

[13] Note: One might try an argument with *mixed modalities*. For instance, one might construe the first premise as (A1'), and the second as (B2'). This approach, however, would render the argument invalid. The inference in the third premise would then essentially be of the of the form: we know *a posteriori* that *p;* for all we know *a priori q*, so for all we know *a priori* (*p and q*). But consider: I know *a posteriori* that the earth is round. And yet for all I know *a priori*, the earth is flat. Does it follow that for all I know *a priori* that the earth is both round *and* flat? No. I know without so much as checking that the earth cannot both be round and flat.

[14] See Lewis (1979) and Kratzer (1977) on relative modality. See also DeRose (1991). Consider the way the relativity can be made explicit with modifying phrases like "in view of q" or "given q":

i. Given the fingerprint analysis, he must be guilty.

ii. Given the total evidence, he must be guilty.

Notice that (i) could be false, even though (ii) is true.

[15] Our *a posteriori* evidence includes any of the relevant "Moorean" evidence, so we won't explicitly mention it.

conscious – after all, there is a conceivable scenario where we are not conscious. The claim instead is that there is no conceivable scenario *where our a posteriori evidence is the same* but we are not conscious. Such a scenario is not ideally conceivable. Premise (2'), then, is correspondingly the claim that it is conceivable that we are merely physical, even if our *a posteriori* evidence is held fixed. That is, there is a conceivable scenario *where our a posteriori evidence is the same* but everything is physical. And now consider the argument disambiguated accordingly:

(B1'*) It is inconceivable that our *a posteriori* evidence holds but we aren't conscious.

(B2'*) It is conceivable that our *a posteriori* evidence holds but everything is physical.

(B3'*) If (B1'*) and (B2'*), then it is conceivable that we are conscious but everything is physical.

(B4'*) If it is conceivable that we are conscious but everything is physical, then it is possible that we are conscious but everything is physical.

(B5'*) So, it is possible that we are conscious but everything is physical.[16]

But understood in this way, though the dualist will (or certainly could) accept (B4'*) (as well as (B1'*) and (B3'*)), now she will simply reject (B2'*). According to the dualist, it is not conceivable that *we* – we who are conscious! – are merely physical. After all, look at the *a priori* arguments that being in a purely physical world precludes being conscious. So, the dualist claims, (B2'*) is false.

But is denying (B2'*) problematically hubristic? Well, why should it be? The stance here isn't the hubristic "I am absolutely certain that no oracle is going to reveal that the world is purely physical." Indeed, the stance needn't be hubristic at all; one can concede that certain kinds of evidence might come in later that would suggest that, in fact, everything is indeed physical – in which case the dualist would have to conclude that the *a priori* arguments (for dualism) had gone wrong somewhere, even if she can't say where. That is, the dualist can be humble by accepting (A2'), and (B2') for that matter, while nevertheless denying (B2'*) – which denial just amounts to standing by her *a priori* arguments. But it is (B2'*) that is required for the success of the argument.

On this diagnosis, the initial appeal of the epistemic argument for physicalism is due to a slide in locutions like "it might turn out that we are merely physical" or "we can't rule out that we are merely physical". Those go down easy when understood as "for all we know with certainty we

[16] Notice that the consequent of (B3'*) drops the relevant conjunct about our *a posteriori* evidence. This omission, however, is harmless: If it is conceivable that our evidence is the same yet we are conscious and merely physical, then, of course, it is conceivable that we are conscious and merely physical. The general form of the argument is as follows:

$\blacksquare(e \rightarrow c), \blacklozenge(e \wedge p), (\blacksquare(e \rightarrow c) \wedge \blacklozenge(e \wedge p)) \rightarrow \blacklozenge(c \wedge p), \blacklozenge(c \wedge p) \rightarrow \Diamond(c \wedge p) \vDash \Diamond(c \wedge p)$.

are merely physical" or even when understood as "for all we know given ideal rational reflection we are merely physical". But for the argument to work, the key premises need to be interpreted with an idealised sort of epistemic possibility, which is also relativised to our *a posteriori* evidence, viz. "for all we know given ideal rational reflection our *a posteriori* evidence holds but we are merely physical" – and here (2') is, at least, much easier to resist.

4. Flip-flopping incompatibilist

Now we can apply these lessons to the epistemic argument for compatibilism. Our contention, unsurprisingly, is that it similarly equivocates. As before, there are at least *two* pertinent disambiguations of the argument. Here we can be brief. Consider first the disambiguation in terms of certainty:

(A1) We are certain that we are free.
(A2) For all we know with certainty, everything is predetermined.
(A3) If (A1) and (A2), then for all we know with certainty we are free but everything is predetermined.
(A4) If for all we know with certainty we are free but everything is predetermined, then it is possible that we are free but everything is predetermined.
(A5) So, it is possible that we are free but everything is predetermined.

But now the problem: the incompatibilist needn't grant (A4). Again: for the relevant premise to be plausible, the epistemic modality must be *ideal* conceivability. Thus, consider the alternative disambiguation (with the requisite relativisation to current evidence):

(B1*) It is inconceivable that our *a posteriori* evidence holds but we aren't free.
(B2*) It is conceivable that our *a posteriori* evidence holds and everything is predetermined.
(B3*) If (B1*) and (B2*), then it is conceivable that we are free but everything is predetermined.
(B4*) If it is conceivable that we are free but everything is predetermined, then it is possible that we are free but everything is predetermined.
(B5*) So, it is possible that we are free but everything is predetermined.

But now the incompatibilist has available a similar reply to the one developed above: (B2*) is false. They contend that there is no conceivable scenario where our *a posteriori* evidence is the

same but everything is predetermined. But this is *not* to say that we know with certainty that not everything is predetermined, given our (*a posteriori*) knowledge of our own freedom. That is to say: the incompatibilist grants (or can grant) that, as far as we know with certainty, we are predetermined, even given the fact that we are free. But this is *not* to concede that this is ideally conceivable; that is to say, the relevant incompatibilist contends that it is not conceivable that *we who are free* are also predetermined. And yet: the incompatibilist can plainly nevertheless grant that this very contention is one about which she might be mistaken: for all of which we can be certain, we are predetermined. But this concession in no way supports the claim that this epistemic possibility is a genuine possibility.

5. Reply to Fischer

We have argued that Lewis' epistemic argument for compatibilism can be resisted, *even if* the incompatibilist grants what we called "the data" underlying the argument – viz., that we know that we are free, and that for all we know (at least in some sense) we are predetermined. The resulting position, however, commits the incompatibilist to what we earlier called *flip-flopping*: if they were provided convincing reason to think that determinism is true, then they'd give up their belief in incompatibilism. Fischer, however, has argued that this suite of attitudes is unstable. It is thus worth considering Fischer's arguments against what he calls van Inwagen's "flip flopping".[17]

Across a wide body of work, one of Fischer's central themes is that the incompatibilist's belief in freedom must be "held hostage" to the (epistemically) possible empirical discovery that determinism is true – and that this is some sort of cost for incompatibilism. Van Inwagen, however, maintains that incompatibilism is true, but that his belief in freedom is *not* held hostage in this way: in the event that the relevant evidence came in, van Inwagen would conclude that his argument for incompatibilism – the Consequence Argument – must have gone wrong somewhere. Importantly, then, van Inwagen maintains that incompatibilism is true, but that his belief in free will can be *resilient* in the face of the epistemic possibility of determinism.

Fischer protests. He begins by picking up on van Inwagen's (1983: 150) statement that, when it comes to the choice between compatibilism and libertarianism, he chooses the "puzzling" (libertarianism) rather than the "inconceivable" (compatibilism). But now Fischer writes as follows:

[17] For one recent examination of Fischer's arguments here, see Bailey and Seymour (2021). We don't disagree with the diagnosis offered by Bailey and Seymour; what we offer below complements that diagnosis.

> But is this right? Is it really inconceivable for van Inwagen that causal determinism is compatible with freedom and moral responsibility? After all, as I've already noted, he has written that, if he were to be convinced of the truth of causal determinism, he would … embrace compatibilism. My question is simple: how then could it be *inconceivable* that compatibilism is true? Perhaps van Inwagen's point is that at the present moment – in the absence of a compelling reason to accept causal determinism – it is inconceivable to him that compatibilism is also true, but that if he were convinced of the truth of causal determinism, it would (under those rather different circumstances) be conceivable to him that compatibilism is true. But this seems strange and a little awkward. If it *would be* conceivable under the envisaged circumstances that compatibilism is true, why isn't it *now* so conceivable? If one believes that under the counterfactual circumstances in question, there would be no barrier to conceiving of the truth of compatibilism, why is there *now* a barrier to conceiving of the truth of compatibilism? The change in circumstances appears to be irrelevant to the *conceivability* of compatibilism. (2016: 52 – 53)

There seems to be something to this complaint. But it is not immediately clear what the problem is. First, consider the counterfactual (as uttered, and endorsed, by Peter van Inwagen):

> (V1) If I were convinced that determinism holds, then I would *not* be convinced that no one is free. (I'd instead flip-flop and accept compatibilism.)

Fisher asks: if van Inwagen accepts (V1), then how can he maintain that compatibilism is nevertheless inconceivable? After all, van Inwagen certainly grants that it is possible that he comes to believe determinism. If counterfactual van Inwagen is convinced that he is both free and predetermined, then of course counterfactual van Inwagen thinks that the compatibilist thesis is coherent – he is embracing it after all! Fischer then insists (in effect) that if it is possible that it is conceivable that *p* then it is indeed conceivable that *p*. And, with a certain understanding of "conceivable", that principle seems plausible enough. So, by accepting (V1), it seems that van Inwagen ends up endorsing the counterfactual possibility, and thus the actuality, of the conceivability of compatibilism.

But here we must be careful. Counterfactual van Inwagen has certain beliefs and says certain things about the truth of compatibilism. But this shouldn't be taken to support the claim that in this counterfactual world compatibilism is indeed rationally *conceivable*. All that follows is that counterfactual van Inwagen cannot detect any contradiction in compatibilism. Perhaps after

some sustained rational scrutiny, and in light of the new scientific evidence, he says "Aha! Compatibilism is ideally conceivable, after all". But there is no guarantee that counterfactual van Inwagen is correct about this, and so there is no guarantee that compatibilism is, in fact, ideally conceivable.[18]

That's a good response to Fisher's complaint, as stated. As we wish to bring out, however, Fischer's points seem to go a bit deeper. First, we want to make a simple observation: considering "counterfactuals" about what van Inwagen would and would not do in certain circumstances is plausibly a distraction as regards Fischer's core complaint. The issue is not so much about features of van Inwagen's beliefs across counterfactual worlds; it instead concerns van Inwagen's *actual* epistemic state and how it fares in light of various hypotheses *considered as actual*.

Here is a comparison. Imagine someone who strongly believes that Shakespeare wrote *Hamlet*. But now imagine that she engages in a conversation with certain conspiracy theorists. She is asked, "What if the historians reveal that Shakespeare didn't write *Hamlet*? What then?". Here she is being asked to consider the relevant possibility *as actual*, not as merely counterfactual. In this case, she could respond in either of the following two ways:

a. If I were convinced that Shakespeare didn't write *Hamlet*, then I'd be convinced that someone else did.
b. If Shakespeare didn't write *Hamlet*, then someone else did.

The first conditional uses the subjunctive mood, and concerns what the speaker *would* accept were she to accept thus-and-such. The latter instead is an indicative conditional which directly expresses the speaker's commitments. In this case, however, the difference is one more of style than substance. Indeed, it seems that what it is for our subject to accept the indicative conditional

[18] Here we have assumed that the counterfactual (V1) has a possible antecedent – that is we have assumed that is possible that van Inwagen is convinced that determinism holds (while retaining his belief in freedom). This is easily confused with what – by the lights of van Inwagen – would be a *counterpossible*. Consider (V1*) "If I were to *learn* that determinism is true, then I would accept compatibilism". Assuming that the worlds under consideration are ones where van Inwagen is still free, the worlds under consideration are ones where free agents are predetermined. According to van Inwagen such worlds are simply impossible. So, the most natural way for van Inwagen to entertain (V1*) is for him to entertain – what is for him – a counterpossible. Counterpossibles are notoriously vexed. But, in any case, it's difficult to see how anything substantial might follow from van Inwagen accepting (V1*). If he were to learn that he is free and predetermined, he would accept compatibilism. But, so what? If we were to learn that 2 isn't prime, we'd be convinced that no even numbers are prime. But nothing follows about the conceivability of *all primes are odd*. An incompatibilist can, of course, accept that in the described impossible world there would, *per impossibile*, be a flaw in the Consequence Argument. But this doesn't even imply that it is *possible* that there is some flaw in the argument, let alone the actuality of a flaw. Cf. Bailey and Seymour (2021). More on counterpossibles below.

(b) just is for her to be such that she would be convinced of its consequent were she to be convinced of its antecedent—that is, for a subject to accept (b) just is for the counterfactual (a) to hold with respect to that subject.

In the background here is a very natural picture: to believe an indicative conditional is to be disposed to accept the consequent on updating with the antecedent.[19] Consider someone who claims to believe that if the lights are on, then Anders is in his office. We find out: the lights are on. And yet this person *does not* conclude that Anders is in his office; instead, the person hems and haws. Isn't this strong reason to conclude that this person *did not* in fact believe the relevant conditional?

This helps to bring out what may be behind at least some of Fischer's puzzlement with flip-flopping. The issue concerns *what it is* to believe incompatibilism in the first place. That is, insofar as van Inwagen accepts incompatibilism, he would seemingly have to accept the following:

(V2) If determinism holds, then no one is free.

But wait. Does van Inwagen really *believe* this conditional – viz, that if determinism is true, then no one is free? Apparently not: after all, van Inwagen seemingly is *not* disposed to accept the claim that no one is free on coming to accept that determinism is true. Indeed, van Inwagen says that (V1) holds: he'd retain his belief in freedom on coming to accept that determinism is true. So how then can van Inwagen genuinely claim even to *believe* that incompatibilism is true? Believing incompatibilism, one might reasonably think, *requires* that one be disposed to reject freedom on coming to accept determinism. But van Inwagen – it seems – has no such disposition. It can thus appear that van Inwagen only gives *lip service* to incompatibilism. He does not in fact *believe* incompatibilism. To summarize:

1. If S believes incompatibilism, then S believes (V2).
2. Van Inwagen does not believe (V2). So,
3. Van Inwagen does not believe incompatibilism.

[19] Cf. Ramsey (1931) and the associated "Ramsey Test", or Mellor (1993): "'If P, Q' . . . expresses a disposition to infer Q from P. In other words, fully to accept a simple 'If P, Q' is to be disposed fully to believe Q if I fully believe P." (236) For a recent development and extension of a view of this kind, see Khoo (2022).

We can respond to this argument by rejecting either (1) or (2). Consider first a rejection of (2). Again, the thought behind (2) is that believing (V2) amounts to being disposed to reject freedom on accepting determinism – but (crucially) van Inwagen has no such disposition. However, this latter contention is perhaps too quick. First, observe that any theory that links belief in conditionals to dispositions to believe will have to accept that certain such dispositions can be *masked* (cf. Lewis 1997). For instance, perhaps the vase is disposed to shatter if dropped. But perhaps it is nevertheless the case that, due to the presence of a certain sorcerer, if it were dropped, it wouldn't shatter. The vase is disposed to shatter on being dropped – and yet, if dropped, it wouldn't shatter, because in the nearest worlds in which it is dropped, it is protected by a sorcerer. Similarly, one might contend that insofar as van Inwagen believes incompatibilism, he *does* have the disposition in question – viz. to deny freedom on accepting determinism – but this disposition is masked by his firmly held conviction that he is free. In particular, van Inwagen is disposed to reject freedom on accepting determinism – and yet, if van Inwagen were to accept determinism, van Inwagen wouldn't reject freedom, because in the nearest worlds in which van Inwagen accepts determinism, van Inwagen disbelieves incompatibilism.

But we can also respond to this argument by simply denying (1).[20] Consider an example. Consider an atheist who accepts, on *a priori* grounds, the standard argument from evil, according to which the existence of God is incompatible with the existence of evil. That is, this atheist is a God/evil incompatibilist: there is no world in which both God exists and evil exists. Since this atheist takes it as obvious *a posteriori* that evil *does* exist, this atheist concludes, of course, that God does not. Now the point. Is this atheist committed to the truth of the following indicative conditional?

(G) If God exists, then there is no evil.

Not obviously. Indeed, if anything, it is obvious that our atheist *will not* be willing to accept (G).[21] That is, the atheist who accepts the standard argument from evil is not likely to think that if God actually does exist, then there is no evil. Instead, this atheist is likely to think that if God actually does exist, then her *a priori* argument from evil is somewhere mistaken, even if she can't say

[20] Cf. the discussion of this issue in Todd (forthcoming).

[21] Cf. Stalnaker (1975), who contends that an indicative conditional 'if p, q' presupposes that p is compatible with the common ground, or that p is epistemically possible. On this account, since (G) presupposes that God might exist, the atheist will not accept (G). Of course, the atheist could *accommodate* the presupposition – but then there is no reason why the atheist couldn't say, "Of course, God doesn't exist, but if I'm wrong about that, and God does exist, then somehow God and evil are compatible after all." In this case, the atheist still does not accept (G).

where. In other words, the atheist is likely to accept that if God actually does exist – has existed this whole time – then *of course* there is still evil; it is just that God and evil are after all somehow compatible. The upshot is as follows. One can rationally accept that there is no possible world in which both *p* and *q,* and yet *not* accept (the indicative conditional) that if *p, ~q*. This result is puzzling, but it is compelling – and it is even more compelling on the theory (noted above) that links belief in an indicative conditional 'if *p*, *q*' to a disposition to infer *q* on accepting *p*. In particular, the relevant atheist probably isn't disposed to conclude that there is no evil on coming to accept that God actually exists, but instead to conclude that the argument from evil was somewhere mistaken.[22] The general point here is the following. One can accept on *a priori* grounds that there is no world in which both *p* and *q*. However, it may nevertheless be the case that one is *not* disposed to reject *q* on finding out that the *actual world* is a *p*-world; instead, finding out that *p* may lead one to reject the incompatibility of *p* and *q*. More generally, the God/evil example plausibly shows that there is *in principle* nothing problematic about believing that there is no world in which both *p* and *q*, and yet *not* accepting the indicative 'if *p*, ~*q*'. And this is what matters.[23]

6. Libertarianism?

Let's take stock. We have argued that the best version of the epistemic argument for compatibilism is the disambiguation in terms of conceivability: (B1*)-(B5*). And we have insisted that the best way to resist this argument, a way we argued is not threatened by the flip-flopping charge, is to deny (B2*) – viz., that it is conceivable that our *a posteriori* evidence holds and everything is predetermined. But now the seemingly simple question: is (B2*) nevertheless *plausible*? One way of approaching this issue is simply to investigate what *denying* (B2*) must look like. And it is here that we can uncover the epistemic argument for compatibilism in another guise. To deny (B2*) (while accepting the other premises) is to adopt the *libertarian* view. This view is seemingly committed to the cogency of the following pattern of reasoning: we are free [look at the *a posteriori* evidence!], freedom requires indeterminism [look at the *a priori* evidence!], so determinism is false. That is, if it is genuinely impossible that our *a*

[22] Of course, one could – as considered previously – insist that the relevant atheist *is* disposed to reject the existence of evil on coming to accept that God exists, but that this disposition is simply masked.
[23] Another example: dualism. Imagine characterizing dualism – i.e., consciousness and "everything is physical" incompatibilism – as follows: "The dualist contends that if we found out [e.g., from an oracle] that everything is physical, we'd have to conclude that there is no pain." Or: "According to the dualist, if everything is physical, there is no pain." Both characterizations are plainly ridiculous. According to the dualist, if everything is physical, she is totally mistaken about the prerequisites of pain.

posteriori evidence holds and that determinism is true, and we believe this, then it seems that we can reason from the presence of the relevant *a posteriori* evidence to the conclusion that determinism is false.

And that can seem objectionable. But why? The complaint is that this is not an argument of the *right kind* for the falsity of determinism (cf. van Inwagen 1983: 210). Importantly, the complaint isn't that libertarians claim to know that determinism is false *a priori*. They don't (or certainly they needn't). But it is nevertheless true (so the complaint goes) that they claim to (be in position to) know that it is false "from the armchair"— or, at least, without putting on a lab coat, or something of the kind. But one might complain that the question of determinism is a question for *physics*, not philosophy. Questions concerning rival interpretations of quantum mechanics remain an ongoing dispute in the foundations of physics. And here, it seems, we simply must await the further progress of science. Which is it? The collapse hypothesis, Bohmian Mechanics, the many-worlds interpretation, or what? These differing theories provide different answers to the question whether or not the physical laws of the universe are deterministic. Thus, the libertarian is seemingly trespassing in some domain in which they shouldn't, prejudging this dispute *in physics* by insisting that the universe is indeterministic.

In other words, on the libertarian picture, reality appears to be, in a certain sense, problematically *Janus-faced*. On the one hand, we can come to discover the fundamental laws of physics *via* the traditional methodology – whatever exactly that is – employed by *physics*. (Indeed, if we discover that the laws of physics are deterministic *via* this sort of methodology, our libertarian flip-flops.) But we can *also* come to discover (something about) those laws from what appears to be (broadly) *a priori* reflection combined with our own introspection or phenomenology (or, as in van Inwagen (1983: 204-7), what is required for "moral responsibility"). But what exactly are the laws of physics such that they could be found out about in such radically different ways? The libertarian thus faces a problem of an *evidential mismatch*: the tools of discovery (introspection, phenomenology…) are not appropriately matched with the thing to be discovered (the nature of the laws of physics).[24]

We might look at the complaint this way. Consider again the following pattern of reasoning: We are free; freedom requires that the laws are indeterministic; so they are. To maintain that we are not entitled to believe the conclusion on the basis of these premises is equivalent to saying that accepting the second premise gives one a defeater for the first, and vice

[24] On this score, it is notable that Lewis – echoing what is surely compatibilist orthodoxy – writes: "Whether our world is governed by indeterministic laws is settled neither by the Moorean fact that we make free choices nor by *a priori* principles of sufficient reason. Rather, it is a contingent question of theoretical physics" (Lewis 2004: 79).

versa. In other words: accepting incompatibilism – that freedom requires indeterminism – in effect gives one a defeater for one's belief in freedom. Once one comes to accept incompatibilism – viz., that freedom requires something metaphysically substantive, as it were – one must (at the least) *withhold* on belief in freedom, since one (so the thought goes) is *not* entitled to now claim to know that indeterminism obtains. Again, one cannot come to know that indeterminism obtains "from the armchair" – i.e., independently of the methods of science.[25]

Libertarian responses to this complaint will take the form of explaining how it is that we indeed *can* know that determinism is false independently of such methods. Here, we want to bring out how there are (at least) two crucially different ways libertarians may try to pursue this project.

First, it is worth considering certain well-known sociological facts associated with libertarianism – viz., its strong connection with *theism*.[26] There is, we should observe, a sort of theistic rejoinder to this evidential mismatch problem. If we have good reason to think that God exists and designed the natural world (and the laws that govern that world) precisely so as to make possible (*inter alia*) human freedom, then if we discover (*a priori*) that indeterminism is required for human freedom, we thereby discover good reason to think that God would have made the universe and its laws suitably indeterministic. In other words, belief in God might, together with other facts, ground rational confidence that the laws of physics are whatever *a priori* reflection says they *need* to be like in order to make possible human freedom. The simple point here is the following: the truth of theism would turn certain questions that may appear to be thoroughly empirical into issues that are at least partly *a priori*.[27] It is worth noting, however, that

---

[25] For a defence of this position, see Balaguer (2010: 137-141). To confront this type of worry, van Inwagen (1983: 210–212) appeals to an analogy with the Moorean response to the skeptic. See Guillon (2014) and Chevarie-Cossette (2021: 203-4) for critical discussion of van Inwagen's position.

[26] In their survey of professional philosophers, Bourget and Chalmers (2014) identified some striking correlations between libertarianism and other philosophical commitments – some of the highest correlations identified in their survey. For instance, libertarianism was correlated with non-physicalism about the mind (.386), theism (.385), and non-naturalism in metaphilosophy (.343). It is fair to say, then, that libertarianism is highly associated (in the popular philosophical consciousness) with theism and non-naturalism. Only 14.6% of philosophers identified as theists, but 50.8% of libertarians identified as theists. And 50% of theists endorse libertarianism, whereas 67% of atheists endorse compatibilism. Further, and importantly, a mere 4.6% of self-described physicalists identified as libertarians, and 69% identified as compatibilists. The results of a further survey from 2021 are in line with these previous results. For example, libertarianism was seen to be strongly correlated with theism, and anti-correlated with both naturalism and physicalism (survey2020.philpeople.org). It is worth noting that it is hard to name more than a handful of confirmed libertarians who are also confirmed non-theists. For one recent explicit defense of libertarianism without theism, however, see Steward (2016).

[27] Balaguer (2010) contends that the question whether we are free in the way required by the libertarian is simply an empirical issue. Perhaps. Our point here, however, is that most actual libertarians will disagree. In other words, why are most (or at least very many) people who identify as libertarians seemingly irrationally confident that we are free in the precise way envisaged by our best libertarian theory? Answer: not because these people feel like they have some special direct insight into the empirical question about whether we

though van Inwagen is a strong theist, van Inwagen (to our knowledge) never appeals to his theism in order to support his libertarianism; our point here is just that – barring appeal to the strategy discussed shortly – he plausibly *needs* to do so. Otherwise he faces the evidential mismatch problem: how can we discover something about the fundamental laws of physics in the relevant way?[28]

But now let us consider a libertarian response that needn't appeal to theism. Observe that, in the first instance, what the libertarian strictly speaking needs is that the laws governing *our own behavior* are indeterministic. This commitment implies something about the laws of *physics* only given a certain further commitment. Indeed, observe that we have implicitly been assuming a thesis that at least some libertarians would wish (or should wish) to call into question. In particular, consider the following:

> **Reduction Thesis**: If the fundamental laws of physics are deterministic, then the laws governing human behaviour are deterministic.

Call the libertarian who denies the Reduction Thesis the *non-reductionist* libertarian. Now, consider a close variant on the pattern of reasoning considered above – viz., one with the second premise made more precise:

> We are free; freedom requires that the laws of fundamental physics are indeterministic; so the laws of fundamental physics are indeterministic.

The *non*-reductionist libertarian is not, she will contend, committed to *this* pattern of reasoning. The non-reductionist will simply reject the second premise: freedom does not require that the laws of fundamental physics are indeterministic, precisely because *the laws governing human behavior* neither are nor supervene on those laws. Instead, the laws governing human behavior are *autonomous* in some relevant way – in much the same way, perhaps, as dualists characteristically claim that consciousness is autonomous with respect to the physical.

---

meet the libertarian's conditions on free will, but instead because, at some deeper level, they do not see this question as exclusively empirical at all: belief in God, and the belief that (likely) God would make possible our free will, turns this apparently wholly empirical question into one that is at least in part *a priori*.

[28] Note: in view of some of the issues to be discussed shortly, we could make similar points about relevant "special science" laws. Certain theists presumably will reason like this: if the laws of chemistry, or biology, or cognitive neuroscience need to be indeterministic in order for there to be free will, then this gives us good reason to think that such laws *are* indeterministic.

Now the point. It seems obvious that our own phenomenal experience (together with *a priori* reflection) does not and could not give us non-trivial insight into the laws of fundamental physics. However, it is vastly more plausible to maintain that our own experience may give us non-trivial insight into the *autonomous* laws governing our own psychology and behaviour. In other words, assuming the Reduction Thesis, there is a kind of evidential "gap" between the relevant levels: phenomenal or introspective evidence is used to conclude something about some set of laws at some *other* level – e.g., the laws of physics. But the non-reductionist, in effect, seeks to deny that there this gap, or "mismatch".[29] Just as the behavior of elementary physical particles (or fields, or whatever) plausibly *directly* bears on the fundamental laws of physics, so similarly the experience of intentional human agents will *directly* bear on the similarly *fundamental* laws governing human behavior. In other words, first-person evidence concerning our own mental states, our phenomenology, our deliberations, and decision-making processes certainly seem *relevant* to generalisations about human behaviour and psychology. In contrast, it is completely unclear how this sort of first-person evidence could bear non-trivially on generalisations about the behaviour of fundamental particles. And since it will likely be granted by all parties that we often have the phenomenal experience *as of there being nothing that determines what we do*, the non-reductionist libertarian claims that it is plausible that indeed nothing *does* determine what we do.[30]

Return, then, to the claim that coming to believe incompatibilism gives one a defeater for one's belief in freedom. Now, given the relevant non-reductionist picture just presented, does one's acceptance of incompatibilism give one a defeater for one's belief that one is free? Of course not. One *already* believed, on independent grounds, that we meet the relevant (indeterministic) conditions for free will (though perhaps not under that description). If one comes to believe that incompatibilism is true on *a priori* grounds, what one has discovered is nothing more than that it is *required* that we meet those conditions in order to *have* free will. But certainly, one is under no pressure, now, to somehow abandon or modify one's belief in freedom. Further, there is no reason to suggest, in this scenario, that one is reasoning as follows: we are free; freedom requires indeterminism; so indeterminism. Instead, one *already* accepted the

---

[29] If one admits that there *is* this kind of gap, then our point above is that one plausibly needs to appeal to theism to close it. In other words: one can either close the gap with theism, or deny the gap with non-reductionism.

[30] Of course, it is disputed whether our phenomenal experience somehow suggests libertarianism; for recent discussion of this issue, see Horgan (2011), Guillon (2014), Deery (2015a), Deery (2015b), Nichols (2015), and Kissel (2018). On this approach, however, it isn't that our phenomenal experience somehow directly supports libertarianism; instead, it supports merely the claim *that we are sometimes undetermined* – and then it is the *a priori* argument that then supports the claim that we *need* to be (in this way) undetermined in order to be free.

conclusion – and one's accepting that conclusion simply *enabled* one to continue believing the first premise on coming to believe the second.[31]

7. Conclusion

Well, where are we? In this paper, we have argued that, holding fixed that we know that we are free, and that *in some sense* we can't rule out the truth of determinism, the best way to resist the epistemic argument for compatibilism is to deny (B2*). And that denial, as we saw, can be developed in at least two competing ways. In the end, it seems, what we have to offer the libertarian is this: either some form of theism, or instead a radical form of non-reductionism. At this stage, of course, some compatibilists may be inclined to say: what this shows is that the epistemic argument is a good one. For we should reject both theism and this kind of non-reductionism – and if that is so, (B2*) stands, and the argument is vindicated. Perhaps. But we see the dialectic at least somewhat differently. What the epistemic argument for compatibilism reveals is something we perhaps knew already – namely, that a libertarian must be prepared to adopt a metaphysics that is, if not explicitly anti-naturalist, certainly out of step with much of what goes under the heading of contemporary naturalism. Whether this cost is a cost that will concern actual libertarians is a matter we must leave for another day.[32]

---

[31] Note: it is clear that any libertarian must regard this argument as *sound*. Thus, when we say that the libertarian needn't "reason as follows", our point, again, is that the libertarian needn't believe the conclusion "on the basis of" her belief in the premises. In this way, the libertarian might claim that the argument displays what has been called "transmission failure" (Wright 1985): the justification or warrant for the premises doesn't transfer to the conclusion. Notably, the reasoning at issue here shares certain features with what has been called "McKinsey-style" reasoning (see McKinsey 1991), where transmission failure is a leading diagnosis. That is, a libertarian seemingly claims to know what their external environment is like (what the natural laws are like) by reflection alone, or merely on the basis of an *a priori* principle and something akin to introspection or phenomenology. For further discussion of McKinsey-style reasoning, see Pryor 2007.

[32] For helpful discussion and/or comments on previous drafts of this paper, we would like to thank Neal Tognazzini, Philip Swenson, Andrew Bailey, Mark Balaguer, Wolfgang Schwarz, Mahrad Almotahari, Derek Ball, Daniel Nolan, and David Chalmers.

References

Alter, Torin. 2007. "On the conditional analysis of phenomenal concepts," *Philosophical Studies* 134 (2): 235 - 253.

Bailey, Andrew M. & Seymour, Amy (2021). "In defense of flip-flopping," *Synthese*. 199 (5-6):13907-13924.

Balaguer, Mark. 2010. *Free Will as an Open Scientific Problem*. Bradford: MIT Press.

Bealer, George. 2002. "Modal Epistemology and the Rationalist Renaissance", in Gendler and Hawthorne, eds., *Conceivability and Possibility*, Oxford: Oxford University Press. Pgs 71–125.

Beebee, Helen, Svedverg, Maria, and Ann Whittle. 2020. "*Nihil Obstat*: Lewis's Compatibilist Account of Abilities," *The Monist* 103: 245 – 261.

Bourget, David & Chalmers, David J. 2014. "What do philosophers believe?" *Philosophical Studies*. 170(3): 465-500.

Braddon-Mitchell, David. 2003. "Qualia and analytical conditionals," *Journal of Philosophy* 100 (3):111-135.

Chalmers, David. 2002. "Does Conceivability Entail Possibility", in Gendler and Hawthorne, eds., *Conceivability and Possibility*, Oxford: Oxford University Press. Pgs. 145–200.

Chalmers, David. 2006. "The Foundations of Two-Dimensional Semantics", in *Two-Dimensional Semantics: Foundations and Applications*, M. Garcia-Carpintero and J. Macia (eds.), Oxford: Oxford University Press, pp. 55–140.

Chalmers, David. 2010. "The Two-Dimensional Argument against Materialism," in *The Character of Consciousness*, New York and Oxford: Oxford University Press.

Chevarie-Cossette, Simon-Pierre. 2021. "Knowing about Responsibility: A Trilemma," *American Philosophical Quarterly* 58: 201 – 216.

Coates, Justin. 2017. "Strawson's Modest Transcendental Argument," *British Journal for the History of Philosophy* 25: 799 – 822.

Daw, Russell and Alter, Torin. 2001. "Free acts and robot cats," *Philosophical Studies* 102: 345-57.

Deery, Oisín. 2015a. "The Fall From Eden: Why Libertarianism Isn't Justified By Experience," *Australasian Journal of Philosophy* 93 (2):319-334.

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

Deery, Oisín. 2021. "Free actions as a natural kind," *Synthese* 198 (1):823-843.

DeRose, Keith. 1991. "Epistemic Possibilities," *The Philosophical Review* 100(4), 581-605.

Fischer, John Martin. 2007. "Compatibilism," In *Four View on Free Will*, ed. M. Vargas. Malden, MA: Blackwell Press. Pp. 44-84.

Fischer, John Martin. 2016. "Libertarianism and the Problem of Flip-flopping," in Timpe and Speak, eds., *Free Will and Theism*. Oxford: Oxford University Press.

Fischer, John Martin and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Frankish, Keith. 2007. "The anti-zombie argument," *Philosophical Quarterly* 57 (229):650–666.

Guillon, Jean-Baptiste. 2014. "Van Inwagen on Introspected Freedom," *Philosophical Studies* 168: 645 – 663.

Hawthorne, John. 2001. "Freedom in context," *Philosophical Studies* 104 (1):63-79.

Hawthorne, John. 2002. "Advice for physicalists," *Philosophical Studies* 109 (1):17-52.

Heller, Mark. 1996. "The mad scientist meets the robot cats: Compatibilism, kinds, and

Counterexamples," *Philosophy and Phenomenological Research* 56 (2):333-37.

Hintikka, Jaakko. 1962. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Ithaca: Cornell University Press.

Horgan, Terry. 2011. "Causal compatibilism about agentive phenomenology." In T. Horgan, M. Sabates, and D. Sosa (Eds.), Supervenience in Mind. Cambridge, MA: MIT Press.

Khoo, Justin (2022). *The Meaning of "If"*. Oxford: Oxford University Press.

Kissel, Andrew. "Indeterministic Intuitions and the Spinozan Strategy," *Mind and Language* 33: 280 – 298.

Kratzer, Angelika. 1977. "What 'must' and 'can' must and can mean", *Linguistics and Philosophy* 1(3):337--355.

Kratzer, Angelika. 1986. "Conditionals", *Chicago Linguistics Society*, 22(2):1–15.

Kripke, Saul. 1980. *Naming and Necessity*. Harvard: Harvard University Press.

Latham, Andrew James. 2019. "The Conceptual Impossibility of Free Will Error Theory", *European Journal of Analytic Philosophy* 15 (2):99-120.

Lewis, David. 1979. "Scorekeeping in a Language Game", *Journal of Philosophical Logic* 8(1): 339-359.

Lewis, David. 1981. "Are We Free to Break the Laws?", *Theoria* 47: 113–121.

Lewis, David. 1986. *On the Plurality of Worlds*. Wiley-Blackwell.

Lewis, David. 1993. "Evil for Freedom's Sake?", *Philosophical Papers* 22: 149–172.

Lewis, David. 1996. "Elusive Knowledge", *Australasian Journal of Philosophy* 74: 549–567.

Lewis, David. 2004. "Causation as Influence," in John Collins, Ned Hall, and L.A. Paul

(eds.), *Causation and Counterfactuals*, Cambridge: MIT Press, pp. 75–106.

Lewis, David. 2020. "Outline of *Nihil Obstat*: An Analysis of Ability", *The Monist* 103: 241- 244.

McKinsey, Michael. 1991. "Anti-individualism and privileged access," *Analysis*, 51: 9–16.

Mele, Alfred. 2019. *Manipulated Agents: A window to moral responsibility*. Oxford: Oxford University Press.

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

Nolan, Daniel. 2015. "Lewis's Philosophical Method," in B. Loewer and J. Schaffer (eds.), *A Companion to Lewis*. Wiley-Blackwell. pp. 25-39.

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

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

Pryor, James. 2007. "What's Wrong with McKinsey-style Reasoning?", in Sanford Goldberg (ed.), *Internalism and Externalism in Semantics and Epistemology*. Oxford University Press, 177-200.

Ramsey, Frank. 1931. "General Propositions and Causality", in *The Foundations of Mathematics and other Logical Essays*. London: Kegan Paul, Trench, Trubner & Co: 237–255.

Schwarz, Wolfgang. 2022. "The Problem of Metaphysical Omniscience", in *Perspectives on the Philosophy of David K. Lewis*, Beebee, Helen & Fisher, Anthony (eds.), Oxford University Press, pp. 23-40.

Speak, Daniel. 2011. "The Consequence Argument Revisited," in R.Kane., ed., *The Oxford Handbook of Free Will*. Oxford: Oxford University Press.

Stalnaker, Robert. 1975. "Indicative Conditionals," *Philosophia* 5: 269–286.

Stanley, Jason. 2004. "On the linguistic basis for contextualism", *Philosophical Studies*: 119(1/2), 119-146.

Steward, Helen. 2016. "Libertarianism as a Naturalistic Position," in Timpe and Speak, eds. *Free*

Will and Theism: Connections, Contingencies, and Concerns. Oxford: Oxford University Press. Pgs. 158 – 171.

Strawson, P. F. 1974. *Freedom and Resentment and Other Essays*. Routledge.

Todd, Patrick. 2017. "Manipulation Arguments and the Freedom to do Otherwise," *Philosophy and Phenomenological Research* 95 (2): 395-407

Todd, Patrick. 2019. "The Replication Argument for Incompatibilism," *Erkenntnis*. 84 (6): 1341-1359. 2019

Todd, Patrick. forthcoming. "The Consequences of Incompatibilism," in *The Routledge Handbook of Responsibility*, ed. M. Kiener.

Van Inwagen, Peter. 1983. *An Essay on Free Will.* Oxford: Oxford University Press.

von Fintel, Kai. 2011. "Conditionals," in von Heusinger, Maienborn, and Portner (eds.) *Semantics: An International Handbook of Meaning*, Vol. 2, 1515–1538. Berlin/Boston: de Gruyter Mouton.

Wright, Crispin. 1985. "Facts and Certainty," *Proceedings of the British Academy* 71: 429–472.