What’s wrong with the consequence argument: In defence of compatibilist libertarianism

Christian List[1]

First version: 25 January 2015; this version: 23 September 2015

1. Introduction

The most prominent argument for the incompatibility of free will and determinism is Peter van Inwagen’s consequence argument (e.g., 1975, 1983, 1989). In this paper, I offer a new diagnosis of what is wrong with this argument. Both proponents and critics of the argument typically accept the way it is framed and only disagree on whether the argument’s premises and the rules of inference on which it relies are true. I suggest that the argument involves a category mistake: it conflates two different levels of description, namely the physical level at which we describe the world from the perspective of fundamental physics and the agential level at which we describe agents and their actions. My diagnosis is based on an account of free will as a higher-level phenomenon that was developed in List (2014).[2] I will call this account ‘compatibilist libertarianism’, for reasons that will become clear below.[3]

[1] C. List, Departments of Government and Philosophy, LSE; autumn 2015: Harvard Law School. This paper can be cited as an online draft. I am very grateful to Jonathan Birch, Robert Kane, and Eddy Nahmias for helpful written comments on this paper, and to Daniel Dennett, Wlodek Rabinowicz, Walter Sinnott-Armstrong, and Laura Valentini for helpful conversations. I have learnt much from my collaboration with Marcus Pivato on a related project (List and Pivato 2015). I also greatly benefitted from conversations with the late Peter Menzies and wish to take this opportunity to express my admiration for Peter’s work and also to refer readers to his own, distinct analysis of the consequence argument (Menzies forthcoming). My work has been supported by a Leverhulme Major Research Fellowship.

[2] Important precursors of this account are Anthony Kenny’s (1978) and Daniel Dennett’s (2003) accounts, which also stress the higher-level nature of free will. I will here, however, use the framework in List (2014), where an explicit formal model of different levels of description is developed; the framework was further extended in List and Pivato (2015). The present paper advances beyond this earlier work by explicitly addressing, and responding to, van Inwagen’s consequence argument. As several critics have pointed out (in correspondence and in discussions), this is an important gap that needs to be filled.

[3] To the best of my knowledge, the label ‘compatibilist libertarianism’ is not yet established in the philosophical literature. I am aware of only one occurrence of the term in a scholarly publication, namely in an article on Locke by Rickless (2000). Some other combinations of compatibilism and libertarianism have been defended under the label ‘libertarian compatibilism’ by Vihvelin (2000) and Arvan (2013). Vihvelin (2013) also explicitly highlights the challenge to explain ‘what’s wrong with the Consequence Argument’ (p. 18) and then develops offers her own distinct response to that challenge. Beebee and Mele (2002) investigate a form of Humean compatibilism (which combines a Humean account of the laws of nature with the thesis that free will is compatible with determinism) and identify similarities between this form of compatibilism and libertarianism. For an overview of the literature on free will, see the handbook edited by Kane (2002), especially Kane’s introduction.

2. The consequence argument

Let me begin with van Inwagen's argument (following the exposition in van Inwagen 1989; see also Vihvelin 2011). At its centre is a modal operator called 'N'. For any proposition p, let Np mean 'p is true, and there is nothing anyone could have done to make it false'. Van Inwagen proposes two rules of inference:

> **Rule Alpha:** From $\Box p$ infer Np, where $\Box$ is an ordinary necessity operator, standing for 'true in all possible worlds'.
>
> **Rule Beta:** From Np and N$(p \rightarrow q)$ infer Nq, where $\rightarrow$ is the material-implication arrow.

Let p_0 be a proposition that describes the fully specified physical state of the world at some time in the remote past. Let l be a proposition that describes the fundamental laws of physics. And let p be a proposition that describes a particular agent's action that we are interested in: an action of which we wish to know whether it was freely performed. The idea is that the action was freely performed *only if* it is *not* true that Np. The argument now goes as follows:

> **Step 1:** $\Box((p_0 \,\&\, l) \rightarrow p)$ (from determinism)
>
> **Step 2:** $\Box(p_0 \rightarrow (l \rightarrow p))$ (from step 1 and logic)
>
> **Step 3:** N$(p_0 \rightarrow (l \rightarrow p))$ (from Rule Alpha)
>
> **Step 4:** Np_0 (a premise)
>
> **Step 5:** N$(l \rightarrow p)$ (from steps 3, 4, and Rule Beta)
>
> **Step 6:** Nl (a premise)
>
> **Step 7:** Np (from steps 5, 6, and Rule Beta)

In short, determinism implies Np, which in turn implies that the action described by p is not free. If we grant van Inwagen's two inference rules, the argument is valid, and the two premises on which it rests – namely Np_0 and Nl – are hard to reject. So, determinism seems incompatible with free will.

3. What can be said in response?

Incompatibilists typically grant some version of the argument and conclude that there could be no free will in a deterministic world. Libertarians further hold that determinism is false and that there is in fact free will. Compatibilists, by contrast, tend to reject the argument. Some offer a compatibilist reinterpretation of the N operator under which Rule Alpha no longer applies. For example, we could adopt a 'conditional' interpretation of an agent's ability, under which Np is interpreted to mean that *if* the agent had attempted to act otherwise, *then* he or she would have succeeded and p would not have been true.[4] This conditional can be true even if, in the actual world, the agent necessarily did not attempt to act otherwise. All that is needed for the truth of the conditional is that its consequent is true (i.e., not-p) in all nearest, albeit counterfactual, worlds in which the antecedent is true (i.e., the agent attempted to act otherwise). And so, Rule Alpha is blocked under the current reinterpretation of Np. Other compatibilists reject Rule Beta, pointing out, for instance, that it would licence some problematic inferences. A further response is to deny that the action can count as free only if Np is false. This response might appeal to those who hold that free will does not require alternative possibilities. Here, however, I will set these familiar compatibilist objections to the argument aside (for a survey, see Vihvelin 2011) and offer a different response.

I will focus on a key feature of the argument whose significance is seldom acknowledged. The argument involves two different kinds of propositions, which include two different kinds of modal notions. It involves, on the one hand, propositions about the fully specified physical state of the world and what it necessitates under the laws of physics and, on the other hand, propositions about the actions an agent could or could not perform. And it combines physical and agential ideas via certain 'mixed' propositions, such Np_0, Nl, and N$(p_0 \rightarrow (l \rightarrow p))$, which place propositions referring to fundamental physics within the scope of the N operator. The argument therefore presupposes that there is a unified level of description at which we can

(i) adequately talk about *both* fundamental physics *and* intentional agency, and

[4] For a recent defence of the conditional interpretation of abilities, see Menzies (forthcoming).

(ii) combine propositions asserting fundamental physical facts with operators capturing agential abilities.

If we did not have such a level of description at our disposal, the argument could not be properly expressed. For the argument to be well formed, we must be able to express all its constituent propositions in a unified language.

From a philosopher's armchair, it is easy to consider this presupposition innocuous or even to miss the fact that there is such a presupposition. The combination of ordinary language and elementary logic in which the argument is standardly formulated seems to allow us to talk seamlessly about everything ranging from elementary particles to human abilities. Yet, I will suggest, the argument's presupposition does not withstand scrutiny. The argument involves a category mistake, illicitly mixing fundamental-physics talk and agency talk.

4. Why the argument's presupposition is problematic

Let us ask what we would need to do to spell out the consequence argument more precisely. We would have to employ scientifically exact language to express each of the propositions occurring in it. Propositions p_0 and l are supposed to describe the full physical state of the world at a particular time and the fundamental laws of physics, and so they would need to be expressed using the resources of our best theory of fundamental physics. Presumably, we would need to use concepts such as elementary particles, fields, and forces, and various equations capturing their dynamics over time. Along with this, the necessity operator $\Box$ would have to express a modal notion suitable for fundamental physics. Up to this point, the language of fundamental physics seems to be the right one.

But now consider proposition p, which is meant to describe a particular agent's action, and the operator N, which is meant to refer to what some agents could or could not have done. Recall that Np means 'p is true, and there is nothing anyone could have done to make it false'. Neither intentional actions nor agents' abilities are things we can talk about in the language of fundamental physics. In that language, we cannot even talk about tables, trees, and chairs – only about particles, fields, forces, and so on.[5] Agency-

[5] As philosophers of chemistry have pointed out, it is questionable whether even simple chemical concepts such as *acidity* can be re-expressed in fundamental physical terms. See, e.g., Manafu (forthcoming).

related concepts, like belief, desire, intention, and choice, are absent from fundamental physics. A sentence such as ‘Christian prefers reading books to watching movies, so he chooses the former over the latter’ does not belong to the language of fundamental physics, to give a simple example.[6]

Consequently, if we wish to talk about agents and their actions, we must switch to a language of psychology, specifically one in which concepts pertaining to intentional agency can be expressed, together with the relevant modal notions: the agential ‘can’.[7] Even the language of neuroscience may be too low-level for that. At best, we may be able to use it to describe the neural correlates of intentional thought and action, but those neural correlates must not be mistaken for the higher-level psychological phenomena they underpin. As many philosophers have argued, we must not confuse the brain with the mind. The brain is a bio-physical system, in which certain neural processes take place. The mind is a higher-level phenomenon, which, plausibly, supervenes on the brain but cannot be identified with it. It is the brain that supports neural processes, and the mind that thinks (for a discussion, see, e.g., Bennett, Dennett, Hacker, and Searle 2007).

It should be clear, then, that fundamental-physics talk and intentional-agency talk operate at two different levels of description. We cannot use the language of fundamental physics to speak of what agents can and cannot do, just as we cannot use the language of psychology, or that of any other special science, to describe the fully specified physical state of the world and the fundamental laws of nature. What is more, each level of description comes with its own modal notions: physical possibility and necessity are not the same as chemical possibility and necessity; and chemical possibility and necessity, in turn, are not the same as biological possibility and necessity, and so on.

We can now observe three points. First, if we tried to formulate the consequence argument in fundamental physical terms, we would not express proposition p and the N operator adequately, because these belong to the agential level. Second, if we tried to formulate the argument in agential-level terms, we would not express propositions p_0 and l as well as the necessity operator $\Box$ adequately, because these belong to the fundamental physical level. And third, it is doubtful whether ‘mixed’ propositions such as Np_0, Nl,

[6] As discussed later, this is not to deny that agency-facts supervene on physical facts.

[7] For a recent discussion of agentive modalities, see also Maier (2015).

N($l \rightarrow p$), and N($p_0 \rightarrow (l \rightarrow p)$) are well-formed at all, because N and p are agential-level expressions, while p_0 and l are physical-level ones. In short, the consequence argument mixes two levels of description that do not go together.

5. The nature of the disconnect between the physical and the agential levels

A critic might object that I am postulating too much of a disconnect between the physical and the agential levels. However, what I am arguing is entirely consistent with the view that everything in the world, including the phenomenon of intentional agency, *supervenes* on the physical. My claim is only that the physical and the agential levels are *conceptually distinct*: we employ a different conceptual repertoire at each of these levels, along with different level-specific modal notions. The picture that I am defending is one of *supervenience without conceptual reducibility*. (For related discussions of the level-specificity of special-science phenomena, see also List and Pivato 2015 and Glynn 2010.)

According to non-reductive physicalism, which I accept for present purposes, the relationship between the physical and the agential levels is the following. Agential properties supervene on physical properties, but are multiply realizable. So, although agential-level facts are completely settled by underlying physical facts, agential-level descriptions are more coarse-grained than physical-level ones. Special sciences such as psychology (but also chemistry, biology, etc.) deliberately abstract away from micro-physical details. They do this for perfectly good scientific reasons, in order to be able to focus on and explain the macro-patterns they are concerned with. An agential property such as a particular person's holding the belief that Obama is the President of the United States or forming the intention to drink a coffee might be realized by numerous different configurations of underlying physical properties and might be equivalent, at most, to an unwieldy disjunction of physical properties. What plays an explanatory role from an agential perspective is the coarse-grained agential property, not its micro-physical realizer. Within the language of physics, we may not even be able to come up with a precise formal expression to capture the 'wild disjunction' of micro-physical properties to which the agential property might correspond. Agential-level descriptions involve concepts that do not map neatly onto corresponding concepts in physics, and vice versa, even though agential-level facts are fully determined by physical ones. (The multiplerealizability point, now widely accepted, goes back to Fodor 1974 and Putnam 1975.[8] For a recent defence of non-reductive physicalism, see List and Menzies 2009.)

6. A simple model

To make the relationship between the physical and the agential levels formally precise, I use a simple model in which the world is represented as a dynamical system (drawing on List 2014).[9] The system is in a particular state at each point in time, and that state may change over time. Let S denote the set of all possible physical states, which are each fully specified and mutually exclusive. Let T denote the set of all points in time, where T is linearly ordered. A *physical history* is a temporal path of the system through its state space, formally a function, denoted h, from T into S, which assigns to each point in time the corresponding state. We can interpret each history as a possible world described at the physical level. Let Ω denote the set of all possible physical histories; this could be either the universal set of all logically possible functions from T into S or, more plausibly, a set consisting of only those functions that are permitted by the laws of physics. Physical-level propositions are, extensionally speaking, subsets of Ω, though of course we normally use sentences in a suitable language to express them.[10] A proposition p is *true* at some history h if and only if h is contained in the relevant subset.

To introduce modal operators such as $\Box$ (necessity) and $\Diamond$ (possibility), we need to define an accessibility relation between the elements of Ω. Whether one history is accessible from another depends on the time in question. Let us say that history h is *accessible* from history h' at time t if and only if the two histories have the same initial

[8] Giving an example from economics, Fodor (1974, p. 103) illustrates the problem as follows: 'I am willing to believe that physics is general in the sense that it implies that any event which consists of a monetary exchange … has a true description in the vocabulary of physics and in virtue of which it falls under the laws of physics. But banal considerations suggest that a description which covers all such events must be wildly disjunctive. Some monetary exchanges involve strings of wampum. Some involve dollar bills. And some involve signing one's name to a check. What are the chances that a disjunction of physical predicates which covers all these events (i.e., a disjunctive predicate which can form the fight hand side of a bridge law of the form "x is a monetary exchange ⇔ ...") expresses a physical natural kind?'

[9] A version of this model, outside the context of free will, can also be found in List and Pivato (2015). For a related formal analysis of multi-level systems, see Butterfield (2012).

[10] An important consequence of this is that the set of *linguistically expressible* propositions may be a proper subset of the set of all *possible* subsets of Ω. If the language is countable (which it typically is), the former set is also countable, while the latter may well be uncountable. As is standard, we define the *conjunction* of two propositions as the intersection of the two sets of histories; their *disjunction* as the union; and the *negation* of a proposition as its complement in Ω.

segment up to time t and diverge, at most, thereafter. Necessity and possibility can now be defined in the standard way. A physical-level proposition p is *necessary* in history h at time t (i.e., '$\Box p$' is true in h at t) if and only if p is true in all histories h' accessible from h at t. Similarly, p is *possible* in history h at time t (i.e., '$\Diamond p$' is true in h at t) if and only if p is true in some history h' accessible from h at t.

So far, we have defined propositions and modal operators at the physical level. To introduce agential-level propositions and modal operators, we need to re-describe our system accordingly. Let $\mathbb{S}$ denote the set of all possible states as described at the agential level. Each state in $\mathbb{S}$ may specify, for instance, the relevant agents' mental attitudes and their actions at the time in question, as well as the state of their environment at a macroscopic level of grain, but not the precise micro-physical configuration of all underlying elementary particles. In line with non-reductive physicalism, I assume that the agential states in $\mathbb{S}$ supervene on the physical states in S, but are multiply realizable, meaning that there exists a many-to-one mapping σ from S into $\mathbb{S}$ which assigns to each physical state the corresponding agential state. Like physical states, different agential states are mutually exclusive. An *agential history* is a temporal path of the system through its agential-level state space. Formally, this is a function from T into $\mathbb{S}$ rather than S, and we now use the notation $\mathbb{h}$ rather than h. Naturally, each physical history h gives rise to a corresponding agential history $\mathbb{h}$. It is obtained by applying the supervenience mapping σ to the given physical history. Formally, we write $\mathbb{h} = \sigma(h)$. Let $\mathbb{\Omega}$ denote the set of all possible agential histories. An agential-level proposition, then, is (extensionally) a subset of $\mathbb{\Omega}$, where the proposition is true at some agential history $\mathbb{h}$ if and only if $\mathbb{h}$ belongs to that subset. Again, we normally use a sentence in a suitable language to express such a proposition.

We define necessity and possibility at the agential level in exact analogy to necessity and possibility at the physical level, using the symbols ⧈ and ◈ instead of $\Box$ and $\Diamond$. An agential history $\mathbb{h}$ is *accessible* from another such history $\mathbb{h}'$ at time t if and only if the two histories have the same initial segment up to time t and diverge, at most, thereafter. An agential-level proposition $\mathbb{p}$ is *necessary* in agential history $\mathbb{h}$ at time t (i.e.,

‘$\mathbb{\Box}\mathbb{p}$’ is true in $\mathbb{h}$ at t) if and only if $\mathbb{p}$ is true in all agential histories $\mathbb{h}'$ accessible from $\mathbb{h}$ at t. Similarly, $\mathbb{p}$ is *possible* in agential history $\mathbb{h}$ at time t (i.e., ‘$\mathbb{\Diamond}\mathbb{p}$’ is true in $\mathbb{h}$ at t) if and only if $\mathbb{p}$ is true in some histories $\mathbb{h}'$ accessible from $\mathbb{h}$ at t.

It should be clear that agential-level propositions, whose extensions are subsets of $\mathbb{\Omega}$, are formally distinct from physical-level propositions, whose extensions are subsets of Ω. Technically, the set of all agential-level propositions, which we may denote $\mathcal{P}(\mathbb{\Omega})$, forms an algebra that is distinct from the algebra of all physical-level propositions, which we may denote $\mathcal{P}(\Omega)$.[11] Agential-level propositions are not contained in $\mathcal{P}(\Omega)$, just as physical-level propositions are not contained in $\mathcal{P}(\mathbb{\Omega})$.[12] Likewise, the modal operators for each of these two levels are distinct, and they range over different domains of propositions. That is, the physical-level modal operators $\Box$ and $\Diamond$ range over the propositions in $\mathcal{P}(\Omega)$, while the agential-level modal operators $\mathbb{\Box}$ and $\mathbb{\Diamond}$ range over the propositions in $\mathcal{P}(\mathbb{\Omega})$.[13]

[11] An *algebra* is a set of propositions that is closed under conjunction (intersection), disjunction (union), and negation (complementation). One might be tempted to define $\mathcal{P}(\Omega)$ and $\mathcal{P}(\mathbb{\Omega})$ as the power sets of Ω and $\mathbb{\Omega}$, respectively (where a set's *power set* is the set of all its subsets). But, as noted, we normally express propositions in some language, and therefore it is more useful to define $\mathcal{P}(\Omega)$ and $\mathcal{P}(\mathbb{\Omega})$ as the sets of propositions that are expressible in, respectively, the appropriate physical-level and agential-level languages. These sets may be smaller than the power sets of Ω and $\mathbb{\Omega}$, respectively (recall footnote 10).

[12] At most, $\mathcal{P}(\mathbb{\Omega})$ might be isomorphic to a sub-algebra of $\mathcal{P}(\Omega)$. This will be the case, in particular, if $\mathcal{P}(\Omega)$ is defined as the power set of Ω. For each $\mathbb{p}$ in $\mathcal{P}(\mathbb{\Omega})$, it will then be the case that $\sigma^{-1}(\mathbb{p})$ is in $\mathcal{P}(\Omega)$, where $\sigma^{-1}(\mathbb{p}) = \{h\in\Omega : \sigma(h)\in\mathbb{p}\}$. However, if we take $\mathcal{P}(\Omega)$ to be the set of all propositions that are expressible in our physical-level language, as suggested earlier, then there is no guarantee that, for every $\mathbb{p}$ in $\mathcal{P}(\mathbb{\Omega})$, $\sigma^{-1}(\mathbb{p})$ is in $\mathcal{P}(\Omega)$. For example, the set of physical-level histories $\sigma^{-1}(\mathbb{p})$ may not be expressible as a finite disjunction in the relevant language. This point was also made in List and Pivato (2015, Section 7).

[13] At each time t, we can think of $\Box$ (respectively, $\Diamond$) as a function that assigns to each proposition p in $\mathcal{P}(\Omega)$ a new proposition $\Box p$ (respectively, $\Diamond p$) in $\mathcal{P}(\Omega)$. Formally, at time t, for any p in $\mathcal{P}(\Omega)$,

$$\Box p = \{h \in \Omega\text{: } p \text{ is true in all histories } h' \text{ accessible from } h \text{ at } t\}\text{, and}$$
$$\Diamond p = \{h \in \Omega\text{: } p \text{ is true in at least one history } h' \text{ accessible from } h \text{ at } t\}.$$

Similarly, at each time t, we can think of $\mathbb{\Box}$ (respectively, $\mathbb{\Diamond}$) as a function that assigns to each proposition $\mathbb{p}$ in $\mathcal{P}(\mathbb{\Omega})$ a new proposition $\mathbb{\Box}\mathbb{p}$ (respectively, $\mathbb{\Diamond}\mathbb{p}$) in $\mathcal{P}(\mathbb{\Omega})$. The definition matches that for $\Box$ (respectively, $\Diamond$), except that we must now replace all lower-level histories h and h', which are elements of Ω, with higher-level histories $\mathbb{h}$ and $\mathbb{h}'$, which are elements of $\mathbb{\Omega}$.

7. Physical-level determinism is compatible with agential-level indeterminism

One feature of this picture of the relationship between the physical and the agential levels is that it renders physical-level determinism compatible with agential-level indeterminism (List 2014).[14] To see how it does this, suppose physical-level determinism is true. Formally, this means that whenever two histories h and h' in the set Ω begin in the same way (i.e., they share some initial state or segment), they must be identical. Figure 1 (reproduced from List 2014) gives an example. It displays six histories over five time periods (from t = 1 in the bottom row to t = 5 in the top row), which make up the set Ω. The dots represent physical states, so that S is the set of all dots. Determinism clearly holds in this case, as any initial segment of any history has only one possible continuation.

Figure 1: World histories at the physical level

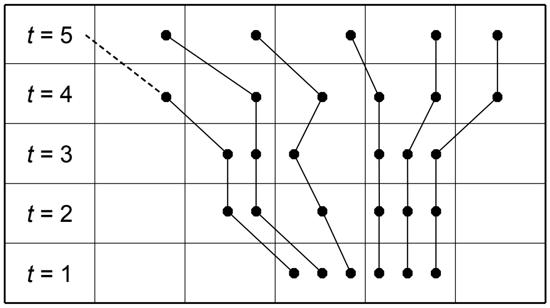
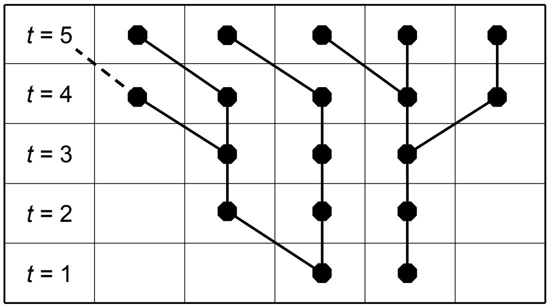


Figure 2: World histories at the agential level



But now suppose that the supervenience relation between physical states and agential states is such that all physical states that fall into the same cell in the rectangular grid give rise to the same agential state. This is an instance of a many-to-one supervenience relation. It implies, for example, that the six physical states at time t = 1 are partitioned into only two distinct agential states, represented by the third and fourth cells from the left. Figure 2 (also reproduced from List 2014) displays the resulting agential histories, which make up the set $\mathbb{\Omega}$. Thick dots represent agential states, so that $\mathbb{S}$ is the set of all thick dots. As we can see in Figure 2, indeterminism is true of agential histories in spite of determinism at the physical level. Two or more agential histories can begin in the same way and then branch off in different directions.

[14] For formally related results outside the context of free will, see Werndl (2009), Butterfield (2012), and List and Pivato (2015).

The same points can be made using the definition of determinism from the consequence argument. Recall that, in that argument, determinism is expressed by the formula $\Box((p_0 \,\&\, l) \rightarrow p)$, where proposition p_0 describes the fully specified physical state of the world at some past time, proposition l describes the laws of physics, and proposition p can, for the moment, be any other physical-level proposition we are interested in. (We thus temporarily set aside the fact that, in the original consequence argument, p has an agential-level interpretation.)

In our example, p_0 is a physical-level proposition that describes the initial state of the system. If the actual physical history is the left-most history in Figure 1 – call it h – then p_0 says that the initial state is given by the left-most dot in the bottom row. Formally, p_0 is the set of all histories that begin in that state, i.e., $p_0 = \{h\}$. Proposition l, which describes the laws, can be taken to be the set Ω in its entirety; we have already built the laws into the specification of Ω. Now it is easy to see that, for *any* physical-level proposition p that is true in history h (i.e., any subset of Ω containing h), we have $\Box((p_0 \,\&\, l) \rightarrow p)$ in history h at any time t. A history in which p is false is simply inaccessible at any time from any history with the initial state described by p_0, under the given laws. So, at the physical level, our system is deterministic according to the definition from the consequence argument.

By contrast, if we re-describe our illustrative system at the agential level and ask whether it is deterministic at that level, we must consider the agential-level version of the relevant formula, namely $\mathbb{\Box}((\mathbb{p}_0 \,\&\, \mathbb{l}) \rightarrow \mathbb{p})$. Here proposition $\mathbb{p}_0$ describes the system's initial state at the agential level, proposition $\mathbb{l}$ describes the system's agential-level laws (which supervene on the physical laws), and proposition $\mathbb{p}$ can be any agential-level proposition we are interested in.

Unlike its physical-level counterpart, $\mathbb{\Box}((\mathbb{p}_0 \,\&\, \mathbb{l}) \rightarrow \mathbb{p})$ is not always true for all true agential-level propositions $\mathbb{p}$. Suppose that the actual history is the left-most one in Figure 2; call it $\mathbb{h}$. Then $\mathbb{p}_0$ says that the initial agential state is given by the left-most thick dot in the bottom row; formally, this is the set of all agential histories that begin in that state (so, $\mathbb{p}_0$ has three elements). In analogy to proposition l above, proposition $\mathbb{l}$ is $\mathbb{\Omega}$ in its entirety; the laws have already been built into our specification of the set of possible histories, and $\mathbb{l}$ (= $\mathbb{\Omega}$) is the projection of the physical-level proposition l (= Ω) under the supervenience relation. To give an example of a true agential-level proposition $\mathbb{p}$ for which $\Box((\mathbb{p}_0 \,\&\, \mathbb{l}) \rightarrow \mathbb{p})$ is false at time 1, simply take $\mathbb{p} = \{\mathbb{h}\}$, a proposition that describes the full truth about the actual history. Since two histories in which $\mathbb{p}$ is false are accessible from history $\mathbb{h}$ at time 1 and both of these histories are contained in the intersection of $\mathbb{p}_0$ and $\mathbb{l}$, we have $\Diamond\sim((\mathbb{p}_0\&\mathbb{l}) \rightarrow \mathbb{p})$ in history $\mathbb{h}$ at time 1 (here '~' stands for 'not'). Thus, under the present definition, our system is indeterministic at the agential level, despite being deterministic at the physical one.

7. Two valid arguments, neither of which establishes the incompatibility of free will and physical-level determinism

Let me return to the consequence argument. If what I have argued is correct, we can identify two valid arguments that are in the vicinity of the original consequence argument, but neither of which establishes the incompatibility of free will and physical-level determinism.

The first argument is formulated entirely at the physical level, where propositions are subsets of Ω, and it looks, on the surface, much like the original argument. As before, propositions p_0 and l describe, respectively, an initial physical state and the fundamental physical laws, and the necessity operator $\Box$ is defined for physical-level propositions. Proposition p, however, cannot literally describe an agent's action, but can at most describe some physical base facts on which the action supervenes, with all the qualifications mentioned above (recall especially footnote 12). The operator N, similarly, is not defined at the physical level and must be replaced by something that can be glossed in physical terms. Accordingly, I will replace '$\mathrm{N}p$' with '$\sim\Diamond\sim p$', i.e., 'not-p is impossible'. This substitution immediately validates both Rule Alpha and Rule Beta. We then obtain the following argument:

Step 1: $\Box((p_0 \,\&\, l) \rightarrow p)$ (from determinism)

Step 2: $\Box(p_0 \rightarrow (l \rightarrow p))$ (from step 1 and logic)

Step 3: $\sim\Diamond\sim(p_0 \rightarrow (l \rightarrow p))$ (from Rule Alpha)

Step 4: $\sim\Diamond\sim p_0$ (a premise)

Step 5: $\sim\Diamond\sim(l \rightarrow p)$ (from steps 3, 4, and Rule Beta)

Step 6: $\sim\Diamond\sim l$ (a premise)

Step 7: $\sim\Diamond\sim p$ (from steps 5, 6, and Rule Beta)

Given the duality of $\Box$ and $\Diamond$, a simpler exposition would of course be possible, without invoking Rules Alpha and Beta, but I have presented the argument in a way that mirrors the structure of the original consequence argument. The argument is clearly valid. However, since p is not the originally intended action-proposition and N is not the originally intended agential modal operator, the argument only establishes a logical relationship between certain physical-level propositions: under the given premises, physical-level determinism renders it impossible for the physical-level proposition p to be false. This hardly establishes the incompatibility of determinism and free will. It only restates a point we have long known, namely that, under physical-level determinism, the initial physical state of the world together with the laws of physics necessitates all subsequent physical states and thereby all physical-level truths about them. Moreover, the argument is formulated at a level at which we could not plausibly expect to find free will. After all, free will is a feature of intentional agents, and intentional agency is a higher-level phenomenon, not a physical one.

The second argument is formulated entirely at the agential level, and it is also structurally similar to the consequence argument, though now propositions are subsets of $\mathbb{\Omega}$ rather than Ω. Here we take propositions $\mathbb{p}_0$ and $\mathbb{l}$ to describe an initial state at the coarse-grained, agential level and the laws governing that level, respectively, and we take the operator $\mathbb{\Box}$ to be defined for agential-level propositions, as explained earlier. Since everything is framed in agential-level terms, we are now able to take $\mathbb{p}$ to be an action-proposition – thereby matching van Inwagen's intended interpretation – and also to interpret N in the intended way. Specifically, we can take 'N$\mathbb{p}$' to mean '$\sim\mathbb{\Diamond}\sim\mathbb{p}$', i.e., 'not-$\mathbb{p}$ is agentially impossible', or more informally: 'there is nothing any agent could have done to render $\mathbb{p}$ false'.[15] Again, Rules Alpha and Beta are valid under this substitution. The argument now runs as follows:

Step 1: $\Box((\mathbb{p}_0 \,\&\, \mathbb{l}) \rightarrow \mathbb{p})$ (from agential-level determinism)

Step 2: $\Box(\mathbb{p}_0 \rightarrow (\mathbb{l} \rightarrow \mathbb{p}))$ (from step 1 and logic)

Step 3: $\sim\Diamond\sim(\mathbb{p}_0 \rightarrow (\mathbb{l} \rightarrow \mathbb{p}))$ (from Rule Alpha)

Step 4: $\sim\Diamond\sim\mathbb{p}_0$ (a premise)

Step 5: $\sim\Diamond\sim(\mathbb{l} \rightarrow \mathbb{p})$ (from steps 3, 4, and Rule Beta)

Step 6: $\sim\Diamond\sim\mathbb{l}$ (a premise)

Step 7: $\sim\Diamond\sim\mathbb{p}$ (from steps 5, 6, and Rule Beta)

This line of reasoning, I think, is impeccable, and it does indeed establish an incompatibilist conclusion of sorts: it establishes that *agential-level* determinism rules out free will. If it turned out that the world was deterministic at the agential level – for instance, because the laws of psychology were such that human agents are deterministic systems – then there could be no such thing as free will (as also conceded in List 2014). But the argument is distinct from the original consequence argument, which purports to show the incompatibility of free will and physical-level determinism.

8. A final objection

We have seen that the consequence argument is well formed only if all its constituent propositions are expressed in the same language, and that neither a *purely* physical-level language nor a *purely* agential-level language is fit for purpose. A physical-level language fails to capture the intended *conclusion* about agential abilities, and an agential-level language fails to capture the intended *premise* about physical-level determinism.

Now, a critic might concede these points and yet insist that the consequence argument can be rescued through the use of some 'mixed-level' language. In particular, the critic might say, cross-level statements are not always ill formed: for example, we

[15] Note further that, on this interpretation and the semantics of $\Diamond$ introduced above, the truth of $\mathbb{p}$ is also a consequence of N$\mathbb{p}$, as required under van Inwagen's intended interpretation of N.

routinely talk about issues such as supervenience, and we thereby express statements about the relationship between properties at different levels. If we can talk about the supervenience of agential properties on physical ones in a 'mixed-level' language, then we should also be able to express the consequence argument in a similar way. Doesn't this show that the argument's mixing of levels can be made to work, after all?[16]

To respond to this objection, we must begin by noting a key desideratum that the mixed-level language would have to fulfil in order to express the consequence argument successfully. The language would have to enable us, not merely to *talk about* the argument from some external ('meta-linguistic') perspective, but to *assert* the argument – i.e., to *use* its constituent propositions, in an 'object-language' way, not just to offer external commentary on them. Arguably, when we engage in supervenience talk, we often adopt an external perspective, for instance by stepping outside any particular level of description and then talking about how what is true at one level relates to what is true at another.[17] For instance, when we consider the level-specific sets of histories Ω and $\mathbb{\Omega}$, together with the associated algebras $\mathcal{P}(\Omega)$ and $\mathcal{P}(\mathbb{\Omega})$, and talk about a function, such as σ, that relates them to one another, we are, in effect, adopting an external perspective. The language in which we are describing the supervenience mapping σ is best understood as referring to the algebras $\mathcal{P}(\Omega)$ and $\mathcal{P}(\mathbb{\Omega})$ from the outside. We are not using that language to *assert* the propositions in those algebras; we are using it only to analyse how they relate to one another. What we need, in order to express the consequence argument in the intended way, is a language in which we can *assert* all of the propositions the argument involves.

One way to construct such a 'mixed-level' language would be to associate it with the smallest algebra containing both $\mathcal{P}(\Omega)$ and $\mathcal{P}(\mathbb{\Omega})$. This algebra is obtained by taking the union of $\mathcal{P}(\Omega)$ and $\mathcal{P}(\mathbb{\Omega})$ and closing it under complementation, conjunction, and disjunction. Clearly, we would then be able to express every proposition from $\mathcal{P}(\Omega)$ and every proposition from $\mathcal{P}(\mathbb{\Omega})$, as well as all logical combinations of these propositions.

[16] Thanks to Jonathan Birch for a very helpful email correspondence that has prompted me to address this objection.

[17] This may even involve attaching explicit 'level indices' – e.g., through notational conventions – to all the propositions we are referring to, so as to indicate the levels to which they belong.

Moreover, since $\mathcal{P}(\Omega)$ contains propositions involving the operators $\Box$ and $\Diamond$, and $\mathcal{P}(\mathbb{\Omega})$ contains propositions involving the operators $⧈$ and $◈$, our new language would be able to express both physical-level and agential-level modal notions. (Note that '$\sim◈\sim$' or equivalently '$⧈$' could be interpreted as 'N'.)

Is this enough to rehabilitate the consequence argument? The answer is 'no', for the following reason. While our mixed-level language would allow us to express physical-level propositions, agential-level propositions, and composite propositions involving conjunctions and disjunctions of the two (and thereby also material conditionals), it would still not allow us to express the kinds of mixed propositions needed for the consequence argument. Recall that the argument involves not just conjunctions or disjunctions of propositions from the two different levels (an example of such a proposition would be p_0 & $⧈\mathbb{p}$); it involves placing physical-level propositions within the scope of an agential-level modal operator, as in Np_0 or, to use this paper's notation, $⧈p_0$. And this is where the category mistake crops up.

Even though our mixed-level language allows us to form conjunctions, disjunctions, and material conditionals involving propositions from both $\mathcal{P}(\Omega)$ and $\mathcal{P}(\mathbb{\Omega})$, the agential-level modal operators $⧈$ and $◈$ occurring in such mixed-level propositions must still range over the propositions from $\mathcal{P}(\mathbb{\Omega})$. So, while a mixed-level proposition such as p_0 & $⧈\mathbb{p}$ is well-formed – being a conjunction of a proposition from $\mathcal{P}(\Omega)$ and a proposition from $\mathcal{P}(\mathbb{\Omega})$ – an expression such as $⧈p_0$ is not.

What is wrong with the latter expression? Its lack of well-formedness lies in the semantics of $⧈$. In brief, p_0 is outside the domain of the operator $⧈$. We can think of that operator as a function defined on $\mathcal{P}(\mathbb{\Omega})$ (as formally explained in footnote 12). It yields well-formed propositions when applied to elements of $\mathcal{P}(\mathbb{\Omega})$, but not when applied to propositions outside $\mathcal{P}(\mathbb{\Omega})$. Thus $⧈\mathbb{p}$ is well formed, while $⧈p_0$ is not.

Let us see what would happen if we tried to truth-evaluate the expression $⧈p_0$. Recall that, for any proposition $\mathbb{p}$ within the scope of $⧈$:

> The expression '$\Box\mathbb{p}$' is true in agential history $\mathbb{h}$ at time t if and only if $\mathbb{p}$ is true in all agential histories $\mathbb{h}'$ accessible from $\mathbb{h}$ at t.

A first difficulty with applying this formula to our mixed-level language is that the *loci* of truth-evaluation for propositions in that language are not *agential-level* histories (which are elements of $\mathbb{\Omega}$), but *physical-level* histories (which are elements of Ω). Only the latter are sufficiently fine-grained to settle all propositions from both $\mathcal{P}(\Omega)$ and $\mathcal{P}(\mathbb{\Omega})$. In order to render the above-quoted formula applicable to our mixed-level case, we would therefore have to stipulate that

> '$\Box p_0$' is true in a *physical-level* history h (at time t) if and only if it is true in the *agential-level* history $\mathbb{h}$ that supervenes on h (i.e., where $\mathbb{h} = \sigma(h)$).

The formula for $\Box$ could then be applied to the right-hand side of this biconditional. However, even if we grant this response to the first difficulty with truth-evaluating expressions such as $\Box p_0$, there is a further difficulty.

Our formula would imply that '$\Box p_0$' is true in an agential-level history $\mathbb{h}$ at time t if and only if

> (*) p_0 is true in all agential histories $\mathbb{h}'$ accessible from $\mathbb{h}$ at t.

But p_0, being a physical-level proposition, is too fine-grained to be settled – in general – by any agential-level history $\mathbb{h}'$. Physical-level propositions such as p_0 do not generally have well-defined truth-values in agential-level histories. An agential-level history such as $\mathbb{h}'$, which belongs to $\mathbb{\Omega}$ rather than Ω, is neither an element of p_0 nor an element of its complement ($\Omega \backslash p_0$), and hence it does not 'decide' between p_0 and its negation (recall that $\sim p_0 = \Omega \backslash p_0$). To illustrate, p_0 may specify some micro-physical properties that are present in some physical realizers of $\mathbb{h}'$ and absent in others. Then p_0 will not have a well-defined truth-value in $\mathbb{h}'$ at all. In consequence, clause (*) is ill-defined, and the entire semantic formula for $\Box$ is not applicable to the expression '$\Box p_0$'. For that formula to yield a well-defined truth-value, the proposition within the scope of the $\Box$ operator must have a determinate truth-value at each agential-level history to which the formula refers. Agential-level propositions meet this requirement; physical-level propositions do not.[18]

To be sure, one could try to respond to this problem by offering a more dramatic redefinition of the semantics of the agential-level modal operators ⧈ and ◈. But this move would be question begging. It would then be unclear whether the redefinition would correctly capture the intended agential-level modal notions. Indeed, if what I have argued is correct, it wouldn't. The agential 'can' is a higher-level notion; it is not to be found at the fine-grained level at which any such redefinition would attempt to relocate it.

9. Concluding remarks

I have argued that the consequence argument involves a category mistake: it conflates two different levels of description, especially by placing physical-level propositions within the scope of agential-level modal operators. We can defend free will against the argument by carefully separating these two different levels of description.

The picture of free will that we end up with is nonetheless incompatibilist in one sense. As I have pointed out, the agential-level variant of the consequence argument establishes the incompatibility of free will with agential-level determinism. This supports a form of 'agential-level incompatibilism'. But crucially – though I cannot defend this point here – there is no reason to think that our best theory of intentional agency will support determinism at the agential level. And so, relative to that level, we have perfectly good grounds for taking free will to be a real phenomenon. Thus a position that we might call 'agential-level libertarianism' is entirely viable.

Furthermore, since indeterminism at the agential level is compatible with determinism at the physical one, the present kind of libertarianism is, in another important sense, a compatibilist position. One might express this point by saying that it is 'intra-level incompatibilist', but 'cross-level compatibilist'. This, I think, justifies calling the resulting view 'compatibilist libertarianism'.

[18] The only physical-level propositions that might conceivably be truth-evaluated in agential histories are ones that are inverse images of sets of agential histories, i.e., propositions of the special form $p = \sigma^{-1}(\mathbb{p})$, where $\mathbb{p}$ is some subset of $\mathbb{\Omega}$. Strictly speaking, however, our truth predicate for physical-level propositions is defined only for subsets of Ω, while that for agential-level propositions is defined only for subsets of $\mathbb{\Omega}$.

Needless to say, more arguments are needed to show that we really do have free will.[19] But the present discussion should help us respond to at least one prominent argument against free will, namely van Inwagen's argument for the incompatibility of free will and physical-level determinism.[20]

References

Arvan, M. (2013) 'A New Theory of Free Will', *Philosophical Forum* 44(1): 1-48.

Beebee, H., and A. Mele (2002) 'Humean Compatibilism', *Mind* 111(442): 201-223.

Bennett, M., D. Dennett, P. Hacker, and J. Searle (2007) *Neuroscience and Philosophy: Brain, Mind, and Language*, New York (Columbia University Press).

Butterfield, J. (2012) 'Laws, Causation and Dynamics at Different Levels', *Interface Focus* 2(1): 101-114.

Dennett, D. (2003) *Freedom Evolves*, London (Penguin).

Fodor, J. (1974) 'Special sciences (or: The disunity of science as a working hypothesis)', *Synthese* 28: 97-115.

Glynn, L. (2010) 'Deterministic Chance', *British Journal for the Philosophy of Science* 61: 51-80.

Kane, R. (ed.) (2002) *The Oxford Handbook of Free Will*, Oxford (Oxford University Press).

Kenny, A. (1978) *Freewill and Responsibility*, London (Routledge).

List, C. (2014) 'Free will, determinism, and the possibility of doing otherwise', *Noûs* 48: 156-178.

List, C., and P. Menzies (2009) 'Non-Reductive Physicalism and the Limits of the Exclusion Principle', *Journal of Philosophy* CVI: 475-502.

List, C., and P. Menzies (forthcoming) 'My brain made me do it: The exclusion argument against free will, and what's wrong with it', in H. Beebee, C. Hitchcock, and H. Price (eds.), *Making a Difference*, Oxford (Oxford University).

List, C., and M. Pivato (2015) 'Emergent Chance', *Philosophical Review* 124(1): 119-152.

[19] For some further ingredients of the defence of free will, see List and Menzies (forthcoming).

[20] Of course, my claim that the consequence argument is ultimately unsuccessful should not be interpreted as diminishing the invaluable role the argument has played in sharpening the debate on free will.

Maier, J. (2015) 'The Agentive Modalities', *Philosophy and Phenomenological Research* XC(1): 113-134.

Manafu, A. (forthcoming) 'A Novel Approach to Emergence in Chemistry', in E. Scerri and L. McIntyre (eds.), *Philosophy of Chemistry: Growth of a New Discipline*, Heidelberg (Springer).

Menzies, P. (forthcoming) 'The Consequence Argument Disarmed: An Interventionist Perspective', in H. Beebee, C. Hitchcock, and H. Price (eds.), *Making a Difference*, Oxford (Oxford University).

Putnam, H. (1975) 'Philosophy and our mental life', in *Mind, Language and Reality*, Cambridge (Cambridge University Press), pp. 291-303.

Rickless, S. C. (2000), 'Locke on the Freedom to Will', *The Locke Newsletter* 31: 43-67.

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 (Clarendon Press).

Van Inwagen, P. (1989) 'When is the Will Free?', *Philosophical Perspectives* 3: 399-422.

Vihvelin, K. (2000) 'Libertarian Compatibilism', *Philosophical Perspectives* 14: 139-166.

Vihvelin, K. (2011) 'Arguments for Incompatibilism', *Stanford Encyclopedia of Philosophy* (Spring 2011 Edition).

Vihvelin, K. (2013) *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*, Oxford (Oxford University Press).

Werndl, C. (2009) 'Are deterministic descriptions and indeterministic descriptions observationally equivalent?' *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 40(3): 232-242.