Free will, determinism, and the possibility to do otherwise

Christian List[1]

14 June 2011

I argue that free will and determinism are compatible, even when we take free will to require the ability to do otherwise and even when we interpret that ability modally, as the possibility to do otherwise, and not just conditionally or dispositionally. My argument draws on a distinction between physical and agential possibility. Although in a deterministic world only one future sequence of events is physically possible for each state of the world, the more coarsely defined state of an agent and his or her environment can be consistent with more than one such sequence, and thus different actions can be "agentially possible". The agential perspective is supported by our best theories of human behaviour, and so we should take it at face value when we refer to what an agent can and cannot do. On the picture I defend, free will is not a physical phenomenon, but a higher-level one on a par with other higher-level phenomena such as agency and intentionality.

1. The problem

A standard argument against the compatibility of free will and determinism is the following.[2]

> **Premise 1:** A necessary condition for someone's action to count as free is that the agent can do otherwise.
>
> **Premise 2:** Determinism implies that the agent cannot do otherwise.
>
> **Conclusion:** Either there are no free actions, or determinism is false (or both).

This argument is often thought to challenge the way we think about our place as agents in the world. If we accept its premises, we have to conclude that we have no free will or live in an indeterministic world. Giving up free will is unattractive because free will is central

[1] I wish to record my intellectual debt to Peter Menzies, who has been a great source of inspiration and advice on the subject of non-reductive physicalism. I am also very grateful to Chlump Chatkupt, Roman Frigg, Alan Hájek, Philip Pettit, Kai Spiekermann and Laura Valentini for helpful comments and discussion.

[2] For an excellent overview of the debate on free will, see the *Oxford Handbook of Free Will*, edited by R. Kane (Oxford: O.U.P., 2002), especially Kane's introduction, on which I here draw. A classic exposition is P. van Inwagen's "The Incompatibility of Free Will and Determinism", *Philosophical Studies* 27, pp. 185-199 (1975). My discussion has also benefited from the debate between J. M. Fischer, R. Kane, D. Pereboom and M. Vargas in *Four Views on Free Will* (Oxford: Blackwell, 2007).

to our self-conception as agents capable of rational deliberation and decision-making, and especially to our attribution of responsibility to each other, which lies at the heart of morality and the law.[3] Accepting indeterminism is not much more attractive either. Although some current physical theories, such as quantum mechanics, seem to support it, it is unclear whether a future grand unified theory of physics will retain this indeterminism, and there is little consensus on whether the quantum indeterminacies permitted by current physical theories are ever "amplified" to a macroscopic level where they could affect human beings and other organisms. But even if they were, a further argument would be needed to dispel the worry that they introduce just randomness into the world, not free will.[4]

To be sure, some people will conclude that free will is an illusion, and they will at most try to rescue the idea of responsibility by offering an account of it not anchored in free will in the conventional sense.[5] Others will argue that freely performed actions are genuine *loci* of indeterminism in the world[6] and perhaps try to relate this indeterminism to the quantum indeterminacies sometimes associated with agential interventions in physical systems.[7] But many people will find both options unsatisfactory, and will therefore wish to re-examine the argument against compatibilism.

One much-discussed possibility is to redefine the notion of free will such that it no longer requires the ability to do otherwise, and thereby to give up Premise 1.[8] Perhaps what matters for free will is not the ability to do otherwise, but that one is the author of one's actions, however this is spelt out. When Martin Luther reaffirmed his criticism of

[3] The relationship between free will and responsibility is a non-trivial matter. Free will might be taken to be (i) necessary for responsibility, (ii) necessary and sufficient, (iii) sufficient, or (iv) neither necessary nor sufficient. I lean towards position (i), but my focus here is on the metaphysical issue of free will, not on the normative issue of responsibility.

[4] Although on standard interpretations quantum mechanics supports indeterminism, this is not the case on all interpretations. On the relationship between modern physics and free will, see D. Hodgson, "Quantum Physics, Consciousness, and Free Will", and R. C. Bishop, "Chaos, Indeterminism, and Free Will", both in the *Oxford Handbook* (op. cit.). For arguments that quantum indeterminacies are irrelevant to free will, see, e.g., T. Honderich, "Determinism as True, Compatibilism and Incompatibilism as False, and the Real Problem", and D. Pereboom, "Living without Free Will: The Case for Hard Incompatibilism", also in the *Oxford Handbook* (op. cit.), and D. Dennett, *Freedom Evolves* (London: Penguin, 2003).

[5] For a discussion of these issues, see M. Vargas, "Revisionism", in *Four Views on Free Will* (op. cit.).

[6] See, e.g., R. Kane's discussion of "Libertarianism" in *Four Views on Free Will* (op. cit.).

[7] See, e.g., Hodgson, "Quantum Physics, Consciousness, and Free Will" (op. cit.), and R. Penrose, *Shadows of the Mind* (Oxford: O. U. P., 1994).

[8] See, e.g., H. Frankfurt's famous argument that the ability to do otherwise is not needed for the kind of free will required for responsibility, in "Alternate Possibilities and Moral Responsibility", *Journal of Philosophy* 66, pp. 829-839 (1969).

the Roman Catholic Church by saying “here I stand; I can do no other”, he was not renouncing free will; he was taking full responsibility for his actions, implying that these were necessitated by his character and motives.[9] But although this might be taken to suggest that free will does not require the ability to do otherwise, Luther need not be interpreted as saying that it was *strictly impossible* for him to do otherwise; it would just have amounted to a betrayal of his values and beliefs. He could not do otherwise *without sacrificing his integrity*. Whether or not we agree with this interpretation – and even if we consider full endorsement of one’s actions a necessary condition for free will – we may still feel that a notion of free will that *never* includes the ability to do otherwise is a rather “watered-down” notion. For this reason, I will here keep Premise 1.[10]

My aim is to offer a new perspective on Premise 2. I will argue that free will, even when understood as requiring the ability to do otherwise, is compatible with determinism. Crucially, I claim that this is so even when the ability to do otherwise is interpreted modally, as the *possibility* to do otherwise, rather than in some weaker conditional or dispositional manner. The key idea is that although determinism implies that only one future sequence of events is *physically possible* given the current fully specified state of the world, the more coarsely defined state of an agent and his or her macroscopic environment can still be consistent with more than one such sequence, and thus different alternative actions can be *possible for the agent*. The notion of agential possibility will be defined and defended in detail below. In particular, I suggest that this notion – and the notion of free will analyzed in terms of it – is no less scientifically respectable than other higher-level notions we routinely employ in intentional explanations, such as beliefs, desires and intentions. However, I will also identify conditions under which new scientific developments might force us to give up this compatibilist view. Notably, these conditions would challenge not only free will, but also our established intentional approach to explaining human behaviour.

[9] R. Kane offers a nice discussion of this example in his introduction to the *Oxford Handbook of Free Will* (op. cit.), referring to D. Dennett, *Elbow Room* (Cambridge: C. U. P., 1984).

[10] Notice that the incompatibilist challenge does not go away even if we weaken Premise 1 to the claim that an agent has free will only if, for at least *some* of his or her actions, he or she can do otherwise.

As with many ideas in philosophy, what I am arguing is motivated by, and implicit in, other works in the literature.[11] But when I was reviewing the literature to prepare my teaching on free will, I found no fully satisfactory articulation of the argument I want to defend, and so I hope this paper will prove useful. The paper should be seen as a broad outline of my proposed argumentative strategy, rather than as a detailed development. It is impossible to do justice to all the nuances of the sophisticated debate on free will within the scope of a single paper.

2. The ability to do otherwise

Since the ability to do otherwise lies at the centre of the present debate, and both premises of the argument against compatibilism refer to it, I must clarify how I will interpret that ability. There are at least three kinds of interpretation on offer, a "traditional conditional", a "new dispositional", and a "modal" one. On a traditional conditional interpretation, to say that the agent can do otherwise is to say that:

> (C) If the agent were to try (or choose) to do otherwise, he or she would succeed in doing so.[12]

On a new dispositional interpretation, it is to say (roughly) that:

> (D) The agent has the disposition to do otherwise when, in appropriate circumstances (to be spelt out further), he or she tries to do otherwise.[13]

On a modal interpretation, finally, it is to say that:

> (M) It is possible (in a sense to be spelt out further) for the agent to do otherwise.[14]

[11] In particular, I have drawn much inspiration from Dennett's writings, especially *Freedom Evolves* (op. cit.) but also C. Taylor and D. Dennett, "Who's Afraid of Determinism? Rethinking Causes and Possibilities", in the *Oxford Handbook of Free Will* (op. cit.), and for some time thought that my view was a variant of Dennett's. But whenever I explained my view to others, the response tended to be that my view differed from what people took Dennett's view to be, and sometimes they were even misled by projecting other aspects of Dennett's philosophy onto my view. I have therefore come to the conclusion that, although I wish to acknowledge and emphasize my debt to Dennett, it is best to develop my view independently.

[12] See, e.g., G. E. Moore, *Ethics* (Oxford: Oxford University Press, 1912), and A. J. Ayer, "Freedom and Necessity", in *Philosophical Essays* (London: Macmillan, 1954).

[13] This definition follows M. Fara, "Masked Abilities and Compatibilism," *Mind* 117, pp. 843-865 (2008). For critical discussions of the family of "new dispositionalist" views, see R. Clarke, "Dispositions, Abilities to Act, and Free Will: The New Dispositionalism", *Mind* 118, pp. 323-351 (2009), and A. Whittle, "Dispositional Abilities", *Philosophers' Imprint* 10, no. 12 (2010).

Which kind of interpretation should we adopt? From a purely logical perspective, a conditional or dispositional interpretation seems much more congenial to a compatibilist defence of free will than a modal one. Consider a conditional interpretation. Even if the agent was always going to do one thing rather than another, it can still be true that in the nearest counterfactual world in which he or she tried to do otherwise, he or she would have succeeded, and thus the truth-conditions of (C) can be met. The fact that determinism may have prevented that counterfactual world from becoming actual is irrelevant to the truth of the conditional, and thus determinism does not rule out the ability to do otherwise on a conditional interpretation. Similarly, consider a dispositional interpretation. Having certain unexercised dispositions is generally taken to be compatible with determinism.[15] Even if determinism prevents a particular disposition from being exercised in a given instance, this does not undermine the presence of that disposition, and thus determinism does not conflict with the truth-conditions of (D) and thereby with the ability to do otherwise, dispositionally interpreted.[16]

A modal interpretation, by contrast, seems to make it much harder to reconcile free will with determinism. If free will requires that it be *possible* for the agent to do otherwise, then it is not obvious how there could be free will in a world in which only one future sequence of events is physically possible. However, I will argue that, initial appearances notwithstanding, the modal approach is the one to take, even from a compatibilist perspective.

For a start, it is widely held that the compatibility of determinism and the ability to do otherwise comes too cheap under a conditional or dispositional interpretation.[17] Take a conditional interpretation. Most of us will agree that if an agent is never psychologically capable of trying to take any action other than a single predestined one, perhaps due to

[14] E.g., S. Hurley supports this kind of interpretation in her paper, "Responsibility, Reason, and Irrelevant Alternatives", *Philosophy and Public Affairs* 28, pp. 205-241 (2000), as mentioned again below.

[15] In his critical discussion, Clarke summarizes the central claim of the new dispositionalism as follows: "Since having unmanifested dispositions is compatible with determinism, having unexercised abilities to act … is likewise compatible." See "Dispositions, Abilities to Act, and Free Will" (op. cit.), p. 323.

[16] Some people may regard the mere fact that determinism does not rule out the ability to do otherwise under a given interpretation as a *reductio ad absurdum* of that interpretation. But to make incompatibility with determinism a *requirement* on any good interpretation of the ability to do otherwise would be to settle the debate about compatibilism by stipulation, rather than by argument.

[17] See, among many others, Hurley, "Responsibility, Reason, and Irrelevant Alternatives" (op. cit.), and Whittle, "Dispositional Abilities" (op. cit.).

some deep psychological obsession, then he or she cannot be said to have the ability to do otherwise. Yet, under a conditional interpretation, such an agent *can* be said to have that ability: it can be true that in the nearest counterfactual world in which he or she tries to do something else – a world that could never materialize – he or she would succeed.[18] Similarly, take a dispositional interpretation. Someone may be severely constrained in what he or she can do under actual, local conditions – and thus unable to act otherwise according to our ordinary way of speaking – and yet have a "masked" disposition to act otherwise, in a more global sense. Think of a pianist who is generally disposed to play a particular Beethoven sonata flawlessly but freezes under the extreme pressure of an audition. In these circumstances, there is little he or she can do to perform better in the audition, despite having the right global disposition. In other words, "[t]he dispositional analyses of abilities … latch on to [a] global sense of ability. But such global abilities to do otherwise do not capture the kind of freedom that is necessary for moral responsibility."[19]

In short, neither the truth-conditions of (C), nor those of (D), capture what we normally mean by the ability to do otherwise in the context of free will. To ascribe such an ability to an agent is to make a modal claim, along the lines of (M), and not just a conditional or dispositional one, such as (C) or (D). Susan Hurley summarizes this point succinctly:

> "The ability to do otherwise entails the *outright possibility* of acting otherwise: it entails that there is a causal possibility of acting otherwise, holding all else constant. A counterfactually conditioned disposition to act otherwise is not the same thing as an outright possibility of acting otherwise… That the former is compatible with determinism does not entail that the latter is."[20]

Apart from the fact that an interpretation of the form (M) seems closer to what we ordinarily mean by saying "the agent can do otherwise" than an interpretation of the forms (C) or (D), the latter also run into certain logical difficulties. A useful test for

[18] Later I return to the challenge that psychological, as opposed to physical, determinism raises for free will.
[19] Whittle, "Dispositional Abilities" (op. cit.), p. 21.
[20] S. Hurley, "Responsibility, Reason, and Irrelevant Alternatives" (op. cit.), pp. 205/206, emphasis original.

whether an interpretation of something is adequate is whether that interpretation can be substituted for its target without changing the meaning too significantly.[21] To see that a conditional or dispositional interpretation of the ability to do otherwise fails this test, notice that the sentences

(1) the agent does not try to do X,

and (2) if the agent does not try to do X, he or she cannot do X,

entail the negation of

(3) the agent can do X,

whereas they do not entail the negation of

(3*) if the agent were to try to do X, he or she would succeed in doing X,

or the negation of

(3**) the agent has the disposition to do X when, in appropriate circumstances, he or she tries to do X.

To the contrary, sentences (1) and (2) seem consistent with (3*), and equally with (3**). There is, at the very least, no obvious contradiction between (1), (2) and (3*), or between (1), (2) and (3**), while (1) and (2) unambiguously imply not-(3). (The actual truth-values of these sentences do not matter for present purposes.) And so (3*) or (3**) cannot be equivalent to (3), contrary to a conditional or dispositional interpretation of the ability to do otherwise.

The combination of these interpretational and logical points leads me to conclude that we should interpret the ability to do otherwise in a modal way, that is, in terms of (M), rather than in a conditional or dispositional one, as in (C) or (D). Of course, a modal interpretation is not free from difficulties either. Great care is needed in specifying the precise notion of “possibility” that renders such an interpretation *both* acceptable *and* compatible with determinism. The challenge, in particular, is to arrive at a notion of possibility that is neither too restrictive, nor too permissive. If it is too restrictive, for

[21] The following argument is adapted from K. Lehrer, “An Empirical Disproof of Determinism”, in *Freedom and Determinism* (New York: Random House, 1966), also discussed in B. Berofsky, “Ifs, Cans, and Free Will: The Issues”, in *Oxford Handbook of Free Will* (op. cit.).

instance by inheriting all the restrictions that determinism imposes on physical possibility, then it seems hard for the possibility to do otherwise to get off the ground in a deterministic world. If it is too permissive, for instance by admitting possibilities ruled out by our scientific understanding of the world, then the claim that the agent can do certain things loses its bite. Much of what follows is therefore devoted to getting the notion of possibility right.

3. The argument against compatibilism revisited

Having clarified how I will interpret the ability to do otherwise, I can restate the initial argument against the compatibility of free will and determinism more precisely.

> **Premise 1:** Free will requires that (at the time of interest) more than one alternative course of action is possible for the agent.
>
> **Premise 2:** Determinism implies that (at the time of interest) only one alternative course of action is possible for the agent.
>
> **Conclusion:** Free will and determinism are incompatible.

As already indicated, I will take Premise 1 to be non-negotiable in the present context. Moreover, the argument is clearly valid, and so the only way to resist its conclusion is to give up Premise 2. At first sight, Premise 2 looks like a correct statement of what determinism implies. Strictly speaking, however, the present wording of Premise 2 is not quite accurate, since determinism is a thesis about *physical possibility*, not about *possibility for the agent*. The following wording is more accurate.

> **Premise 2*:** Determinism implies that (at any given time) only one future sequence of events is physically possible.

Crucially:

> **Observation:** The incompatibilist conclusion follows from premises 1 and 2*
> *only if* Premise 2* implies Premise 2.

And I will argue that:

> **Claim:** Premise 2* does not imply Premise 2.

In particular, my claim is that Premise 2* follows from the definition of determinism, while Premise 2 does not. Let me give a rough sketch of my argument, before presenting a more technical version.

Why Premise 2 follows from the definition of determinism*

When we are interested in whether a particular sequence of events is physically possible at a given time, it is appropriate to ask whether that sequence of events is consistent, under the laws of physics, with the full physical description of the world at that time. If determinism is true, then each physical state of the world is consistent with only one future sequence of events. This is what Premise 2* says.

Why Premise 2 does not follow from the definition of determinism

When we are interested in whether a particular action is *possible for an agent*, by contrast, the appropriate frame of reference is not the one given by fundamental physics, but rather the one given by our best theory of human behaviour. Thus the description of the world that matters here is not a (microscopic) physical one, but a (macroscopic) psychological one. Candidate theories that provide the right level of description include some advanced versions of behavioural decision theory, which are currently our best attempts to make sense of intentional agency. In fact, even folk psychology, their more rudimentary counterpart, outperforms physics or neuroscience when it comes to understanding and explaining human behaviour across different domains and outside isolated laboratory conditions.[22]

Let me introduce the term “agential state” to denote the state of an agent and his or her macroscopic environment *as specified by the relevant higher-level theory of human behaviour*. There are various ways of making this definition more precise; it may sometimes be useful, for example, to represent the agential state explicitly as a pair consisting of the agent’s intrinsic state and the state of the environment.[23] But for the

[22] The fact that we have some basic understanding of certain neuro-physiological mechanisms in a limited number of domains hardly challenges this general point.

[23] As a by-product, this representation opens up useful conceptual resources for dealing with Frankfurt’s much-discussed counterexamples to Premise 1, in “Alternate Possibilities and Moral Responsibility” (op. cit.). By distinguishing between (i) an agent’s intrinsic state and (ii) the state of his or her environment, we are in a position to distinguish between cases in which an agent’s inability to do otherwise is due to (i) and cases in which it is due to (ii). This is related in spirit to the strategy that new dispositionalists such as Fara,

general purposes of this paper, I can set these details aside. What matters is that an agential state is more coarse-grained than the fully specified physical state of the world. The agential state *supervenes* on the physical state – it is determined by the physical state and could not vary without variations in it – but it is *multiply realizable*, in the sense that there is typically more than one physical state that gives rise to the same agential state. It is precisely this coarse-grained nature of someone's agential state that can render more than one sequence of events consistent with it.

> **Observation:** An agential state is consistent with every sequence of events that is supported by at least one of its physical realizations.

As long as some of these sequences correspond to different courses of action, it follows that more than one course of action is *possible for the agent*, contrary to Premise 2.[24]

While this first sketch of my argument should convey the basic idea, I will now develop the argument more carefully. In particular, I will offer two mutually reinforcing lines of reasoning, a "bottom-up" and a "top-down" one. Each of them should individually support my view, but together, I suggest, they form a compelling package.

4. A bottom-up argument for the ability to do otherwise

To see why determinism at the physical level does not rule out the possibility to do otherwise at the agential level, it is useful to introduce a simple toy model of the relationship between the lower, physical and the higher, agential level. In this model, the world is represented as a dynamic system, which can be in a number of different states and whose state changes over time.

The set of all possible states of the system is called its *state space* and denoted *S*. We are interested in how the system's state evolves over time. Let *T* denote the set of all

in "Masked Abilities and Compatibilism" (op. cit.), tend to use to respond to Frankfurt. I am grateful to Peter Menzies for drawing my attention to this point.

[24] Relating this to Dennett's terminology in *Freedom Evolves* (op. cit.), one might say that determinism (at the physical level) does not imply inevitability (at the agential one). But Dennett does not offer a precise modal analysis of his notion of "evitability". Taylor and Dennett, in "Who's Afraid of Determinism?" (op. cit.), argue that when we analyse modal claims such as "it is possible for the agent to do otherwise", we should use a "broad method" for selecting the set of possible worlds we quantify over, allowing some wiggle room around the actual world, rather than a "narrow method". This allows us to say that it was possible for the agent to do otherwise, even in a deterministic world, but one might find the approach somewhat *ad hoc*. The modal account I am offering in this paper is intended to be more systematic.

points in time. For example, T could be the set of all natural numbers (an example of discrete time) or alternatively the set of all non-negative real numbers (an example of continuous time). For present purposes, we only need to assume that T is linearly ordered, so that, for any two points t and t' in T, we can say which is earlier and which is later.[25]

A *world history* is a temporal path of the system through its state space. It is formally represented by a function h from the set of time points T into the system's state space S, which assigns to each point in time, t in T, a corresponding state of the system $h(t)$ in S. Let us write Ω to denote the set of all world histories that are deemed possible, for example on the basis of the underlying laws of physics. Different laws of physics give rise to different specifications of Ω.

Subsets of Ω can be interpreted as *propositions*. A proposition p is said to be *true* in all those histories h that are contained in it (recall that p is a subset of Ω), and *false* in all others. This is just the standard extensional way of defining propositions, applied to world histories. Propositions can express, for example, contents such as "at a particular time, or during a particular period, the system is in a particular state, or passes through a particular region of the state space".

Sometimes we wish to refer not to an entire world history, but only to a truncated part of that history, up to a given point in time t. To do so, we introduce the notation h_t to denote the restriction of the history h to the set of all points in time up to t.[26]

With these concepts in place, we can give formal definitions of determinism and indeterminism. Determinism is the thesis that:

> **Determinism:** For any two histories h, h' in Ω and any point in time t in T, if $h_t = h'_t$, then $h = h'$.

That is, any two histories that coincide up to time t are identical in their entirety; so the world history up to any point in time t fully determines its continuation. Indeterminism is the negation of this thesis.

> **Indeterminism:** There exist some histories h, h' in Ω and some point in time t in T such that $h_t = h'_t$ but $h \neq h'$.

[25] The model does not require us to take a view on whether there was a first moment in time.

[26] More generally, we could write $h_{T'}$ to denote the restriction of the history h to some subset T' of T. Then h_t as defined in the main text is simply the special case where T' is the set of all points in time up to t.

That is, there can be two or more distinct world histories that coincide up to some point in time *t* but subsequently branch out in different directions.

In order to analyze the implications of determinism or indeterminism for various modal statements, such as "it is possible that *p*" or "it is necessary that *p*", we need to define the semantics of modal statements. To do so, I will draw on the standard possible-worlds semantics of modal logic and apply it to world histories.[27]

The central notion that we require is that of an *accessibility relation* between world histories, defined as a binary relation *R* on Ω. For any two histories *h* and *h'* in Ω, *hRh*' is interpreted to mean that *h*' is accessible from *h*. Crucially, the accessibility relation depends on our temporal reference point: whether or not one history is accessible from another depends on where in time we are. To indicate this dependency, we append the subscript *t* to *R*, interpreting R_t as the accessibility relation between world histories at time *t*. In particular, a history *h*' is accessible from a history *h* at time *t* if and only if *h*' is a continuation of h_t, the truncated part of *h* up to time *t*. Formally:

> **Accessibility:** For any histories *h*, *h*' in Ω and any point in time *t* in *T*, hR_th'
> if and only if $h'_t = h_t$.

Thus the histories that are accessible from a given history at time *t* are all the different possible histories that coincide with the given history up to time *t* but then branch out in one way or another. It follows that, under determinism, the only history accessible from any history at any time is that history itself. Under indeterminism, two or more distinct histories can be accessible from the same history at a particular point in time. Now the truth conditions of modal statements can be defined straightforwardly.

> **Truth-conditions of modal statements:** "It is possible that *p*" is true in history *h* at time *t* if and only if *p* is true in *some* history *h*' that is accessible from *h* at time *t*. "It is necessary that *p*" is true in history *h* at time *t* if and only if *p* is true in *every* history *h*' that is accessible from *h* at time *t*.

Under determinism, the fact that only one history is accessible from any given history at any time implies that two statements of the form "it is possible that *p*" and "it is possible

[27] I deliberately keep the present model simple. More sophisticated models, based on various temporal logics, are certainly possible. For a helpful introduction to non-classical logics, see G. Priest, *An Introduction to Non-Classical Logic*, 2nd edition (Cambridge: Cambridge University Press, 2008).

that not *p*” can never be simultaneously true. Under indeterminism, by contrast, two such statements can be simultaneously true, since two or more distinct histories can be accessible from the same history at a particular time.

If we interpret the set S in the model as the set of all *physically possible* states of the world, and the set Ω as the set of all *physically possible* world histories, then the model just confirms Premise 2*: if determinism is true, then only one future continuation of the world history is *physically possible* at any time. However, as I have suggested and will defend further in the next section, this does not settle the question of what actions an agent can and cannot do. This is because the appropriate frame of reference for asking whether a particular action is *possible for an agent* is not the one given by fundamental physics, but the one given by our best theory of human behaviour, and such a theory employs a more coarse-grained state space than the physical one.

Let $\mathbb{S}$ be the set of all possible *agential states*, that is, the set of all possible states of the relevant agents and their macroscopic environment *as specified by our best higher-level theory of human behaviour*. (As noted above, we could define each state in $\mathbb{S}$ to have some further structure – in the case of a single agent, for instance, to take the form of a pair consisting of the intrinsic state of that agent and a state of the environment[28] – but my main argument does not depend on this.) In what follows, I will use outlined (“coarsened”) letters to refer to the agential level, while keeping the original letters to refer to the physical level. The states in $\mathbb{S}$ supervene on those in S, but are multiply realizable by them.

> **Supervenience and multiple realizability:** There exists a (many-to-one) mapping σ from S into $\mathbb{S}$ such that each physical state s in S determines a corresponding agential state $\sigma(s)$ in $\mathbb{S}$, but the same agential state $\mathbb{s}$ in $\mathbb{S}$ may be realized by more than one physical state s in S.

Given the mapping σ from physical to agential states, any world history h at the physical level determines a corresponding world history $\mathbb{h}$ at the agential level, where $\mathbb{h}$ is a

[28] This would allow us, for example, to consider equivalence classes of states that coincide in one of the two components, which, in turn, would be useful for distinguishing between internal and external impediments to the agent’s free choice. The approach can be generalized to multiple agents, by representing the agential state in terms of a suitable vector of agent-specific characteristics.

function from the set of time points T into the agential state space $\mathbb{S}$. Specifically, at any point in time t, the agential state $\mathbb{h}(t)$ is determined by applying the mapping σ to the underlying physical state $h(t)$. The supervenience mapping σ, initially defined as a mapping from physical to agential states, thus also yields a mapping from physical to agential histories.

> **The relationship between physical and agential histories:** For any physical history h in Ω, the corresponding agential history is $\sigma(h) = \mathbb{h}$, where, for any point in time t, the agential state is $\mathbb{h}(t) = \sigma(h(t))$.

Let $\mathbb{\Omega}$ denote the set of all possible agential histories thus determined. It consists of every agential history $\mathbb{h}$ determined by some physical history h in Ω via the supervenience mapping σ. So $\mathbb{\Omega}$ is the projection of Ω under that mapping (that is, $\mathbb{\Omega} = \sigma(\Omega)$). As in the case of the original state space Ω, subsets of the agential state space $\mathbb{\Omega}$ can be interpreted as propositions, this time about agents and their actions. A proposition $\mathbb{p}$, here a subset of $\mathbb{\Omega}$, is said to be *true* in all those histories $\mathbb{h}$ contained in it, and *false* in all others.

Re-using the earlier definitions, we are able to analyze modal statements at the agential level. Just as we defined an accessibility relation R_t between world histories at the physical level at any time t, we can define an accessibility relation $\mathbb{R}_t$ between world histories at the agential level.

> **Agential accessibility:** For any histories $\mathbb{h}$, $\mathbb{h}'$ in $\mathbb{\Omega}$ and any point in time t in T, $\mathbb{h}\mathbb{R}_t\mathbb{h}'$ if and only if $\mathbb{h}_t = \mathbb{h}'_t$.

As before, the notation $\mathbb{h}_t$ stands for the restriction of the history $\mathbb{h}$ to the set of all points in time up to t. So, at time t, a history $\mathbb{h}'$ defined at the agential level is *accessible* from another history $\mathbb{h}$ if and only if $\mathbb{h}'$ is a continuation of $\mathbb{h}_t$. Now modal statements at the agential level have the following truth-conditions.

> **Truth-conditions of modal statements at the agential level:** "It is (agentially) possible that $\mathbb{p}$" is true in history $\mathbb{h}$ at time t if and only if $\mathbb{p}$ is true in *some* history $\mathbb{h}'$ that is agentially accessible from $\mathbb{h}$ at time t. "It is (agentially) necessary that $\mathbb{p}$" is true in history $\mathbb{h}$ at time t if and only if $\mathbb{p}$ is true in *every* history $\mathbb{h}'$ that is agentially accessible from $\mathbb{h}$ at time t.

My main point follows immediately. Given the multiple realizability of agential states by physical states, it is perfectly possible for the many-to-one mapping σ from S into $\mathbb{S}$ to be such that determinism at the physical level is consistent with indeterminism at the agential level. While any physical history (in Ω) may have only one possible continuation at any time, namely the history itself, there can be two or more distinct agential histories (in $\mathbb{\Omega}$) that coincide up to time t but then branch out in different directions.

Figure 1: World histories at the physical level

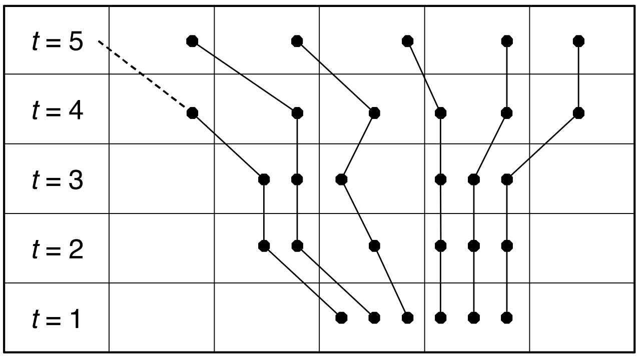
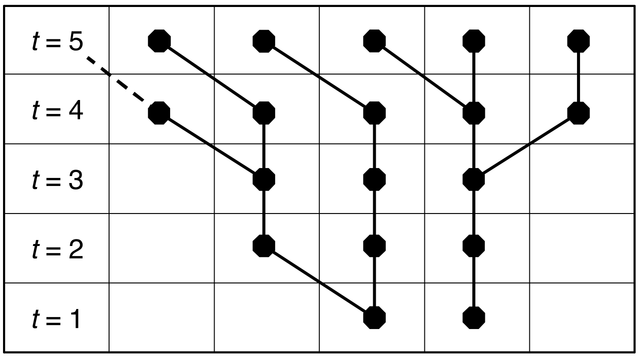


Figure 2: World histories at the agential level



Figures 1 and 2 show a simple example. Figure 1 represents the physical level, Figure 2 the agential one. The dots in Figure 1 represent different possible physical states, and the lines connecting them different possible physical histories, over five time periods (t = 1 to t = 5). Thus S is the set of all the dots, and Ω the set of all the lines. It is easy to see that determinism holds: there is no branching in any of the possible world histories. Now suppose the agential state of the world supervenes on the physical one and is multiply realizable. Specifically, all physical states that lie inside the same rectangular cell in Figure 1 correspond to the same agential state. At time 1, for instance, the supervenience mapping σ maps the three left-most physical states to one agential state, and the three right-most physical states to another. Figure 2 displays all the resulting agential states and world histories. Here the thick dots constitute the set $\mathbb{S}$ of possible agential states, and the thick lines (to be precise, all the linear paths that can be taken along the trees from bottom to top) constitute the set $\mathbb{\Omega}$ of all possible agential histories. It is easy to see that there is no determinism at the agential level in the present example. The agential histories branch out in various ways. As this example shows, such agential-level indeterminism is entirely consistent with determinism at the physical level, jointly with supervenience and multiple realizability.

Now let $\mathbb{h}$, $\mathbb{h}$' be two agential histories (in $\mathbb{\Omega}$) that coincide up to time *t* but then branch out in different directions, and let $\mathbb{p}$ (a subset of $\mathbb{\Omega}$) be a proposition that is true in history $\mathbb{h}$ but not in history $\mathbb{h}$'. The content of $\mathbb{p}$ could be, for instance, that a particular agent performs some action X, where X is something the agent does after time *t* in history $\mathbb{h}$ but not in history $\mathbb{h}$'. Then the statements "it is (agentially) possible that $\mathbb{p}$ (i.e., the agent does X)" and "it is (agentially) possible that not $\mathbb{p}$ (i.e., the agent does not do X)" will both be true in history $\mathbb{h}$ (as well as in history $\mathbb{h}$') at time *t*, and this is precisely the phenomenon that we normally call the *possibility to do otherwise*. The point can be summarized as follows.

> **Proposition:** Given a suitable many-to-one supervenience mapping σ, determinism at the physical level (where the set of possible world histories is Ω) is compatible with indeterminism at the agential level (where the set of possible histories is $\mathbb{\Omega} = \sigma(\Omega)$) and with the agential possibility to do otherwise.

In summary, my simple toy model shows that determinism at the physical level does not rule out the possibility to do otherwise at the agential level, contrary to Premise 2, provided the physical and the agential level are related by a suitable many-to-one supervenience mapping. I now want to defend this picture from a broader philosophical and scientific perspective.

5. A top-down argument for the ability to do otherwise

How should we answer questions about what entities and properties there are in the world and whether something is or is not possible? Although there are many different approaches towards such questions, I will here adopt what is sometimes called the "naturalistic ontological attitude". It can be roughly stated as follows.

> **Naturalistic ontological attitude:** Our best guide to ontological questions in any domain is given by our best scientific theories of that domain.[29]

For example, to determine whether magnetic fields or electrons exist, we must consult our best theories of physics. If these entities are included in the ontological inventory of those theories, we have every reason to consider them real; if not, we don't. The naturalistic approach commands that we take the ontological commitments of our best scientific theories in any given domain at face value. Our ontology of the world will then be contingent on our best scientific theories, and we may sometimes have to revise it in light of new scientific discoveries. Although this will not satisfy everyone's philosophical intuitions, the idea is simply that we have no better guide than science in relation to what entities and properties there are.

It should be evident that the naturalistic ontological attitude by itself does not establish free will and the possibility to do otherwise. To the contrary, if we held the view that all of science is ultimately reducible to physics, and that ontological questions can only be answered at the physical level, then we would have to conclude that there is no room for free will and the possibility to do otherwise in a deterministic world. Neither would be among the ontological commitments of a deterministic theory of physics. At this point, however, my second assumption comes into play.

> **Non-reductive physicalism:** Although the entities and properties studied in the special sciences supervene on physical entities and properties, they are (i) distinct from those physical entities and properties, since they are multiply realizable by them, and (ii) indispensable in causal explanations in the special sciences.[30]

[29] This version is inspired by W. V. Quine's naturalistic realism. See, e.g., *Ontological Relativity and Other Essays* (New York: Columbia University Press, 1977). See also A. Fine, "The Natural Ontological Attitude", in J. Leplin (ed.), *Philosophy of Science* (Berkeley: University of California Press, 1984).

[30] The multiple-realizability claim may be less obvious in the case of entities than in the case of properties. It seems that, at any given time, a special-science entity will be instantiated by one and only one configuration of physical entities. However, the identity conditions of the special-science entity, both across time and across possible worlds, need not match up neatly with the standard identity conditions of its physical realizors. Think of a social-science entity such as a particular firm, a government, or a state. See, e.g., C. List and P. Pettit, *Group Agency* (Oxford: O.U.P., 2011). For a general discussion and defence of non-reductive physicalism, see C. List and P. Menzies, "Non-reductive physicalism and the limits of the exclusion principle," *Journal of Philosophy* CVI (9) (2009), pp. 475-502.

Examples of such special-science entities and properties are an agent's mental states and their possession, on which our psychological ontology of beliefs, desires and intentions is based. Although these supervene on physical entities and properties, such as the relevant organism's brain states, they are multiply realizable in myriad ways and thus distinct from their physical realizors. Moreover, they are indispensable in our current best scientific explanations of human behaviour.[31] Explaining and predicting even basic human interactions at a physical or neuroscientific level seems completely infeasible, whereas an intentional approach as simple as folk psychology has little difficulty making sense of them. Even when we try to understand and explain the behaviour of non-human animals, the ascription of certain intentional states seems indispensable. Behavioural ecologists increasingly make use of the models of decision theory and game theory to explain and predict the behaviour of non-human animals,[32] and any dog owner will know that taking an intentional stance towards a dog is a much more useful guide to its behaviour than trying to understand it as a bio-physical automaton.[33]

If we combine non-reductive physicalism with a naturalistic ontological attitude, it follows that we must consider the entities and properties referred to in our best special-science explanations as real, at least when we talk about the domains of the special sciences. So the beliefs, desires and other properties that our best theories of agency ascribe to individuals have a legitimate place in our ontology. But what about free will and the possibility to do otherwise? Here comes my final premise.

> **Free will as a special-science phenomenon:** Free will, in the technical sense of an agent's having a choice between more than one course of action in many situations, is a key presupposition of our best scientific theories of agency, at least when these theories are understood literally.

As mentioned earlier, I assume that our best theories of agency are some advanced versions of behavioural decision theory, which, in turn, might be seen as vastly improved extensions of folk psychology. Just as folk psychology ascribes free will to people, in that people are assumed to be able to choose from more than one course of action, so our

[31] For some relevant formal results, see List and Menzies, "Non-reductive physicalism" (op. cit.).

[32] For a survey of some relevant works, see L. Conradt and C. List, "Group decisions in humans and animals: a survey," *Philosophical Transactions of the Royal Society* B 364 (2009), pp. 719-742.

[33] On the "intentional stance", see D. Dennett, *The Intentional Stance* (Cambridge/MA: MIT Press, 1987).

more sophisticated theories of agency are committed to free will, at least when they are interpreted literally. A central concept of any version of decision or game theory, whether we take the original versions of von Neumann and Morgenstern, Nash, and Savage or their latest, psychologically more advanced incarnations, is an agent's set of possible actions or strategies. The decision-theoretic or game-theoretic explanation of many social phenomena relies crucially on the assumption that the agents' action-or-strategy sets contain more than one option. Sometimes the addition or removal of options can make a significant difference to what the agents are predicted to do even if these options are not ultimately chosen. Unless we accept that there is at least a thin, technical sense in which such options could have been chosen, it is hard to make sense of those effects.[34]

Of course, in analogy to the instrumentalist view in the philosophy of science, according to which unobservables such as electrons and magnetic fields are just instrumentally useful constructions, one might argue that the ascription of a non-singleton option-set to an agent is just an instrumentally useful way of making sense of that agent's observable behaviour but has no ontological significance. But if we hold a naturalistic ontological attitude, this instrumentalist view is not available to us. To the extent that free will, in the sense of being able to choose from more than one option, is explanatorily indispensable in our best scientific theories of agency, we have to take it at face value.

6. A form of determinism that would threaten the ability to do otherwise

Since my argument for the compatibility of free will and determinism rests on some contingent premises about the nature of our best theories of agency and their non-reductive relationship to physics, it is worth asking under what conditions we might have to abandon the present compatibilist view. It should be clear that we would have to do so

[34] In his working paper, "Free Will: A Rational Illusion" (Tel-Aviv University, 2007), the economist Itzhak Gilboa recognizes rational choice theory's implicit commitment to free will, yet struggles to reconcile this commitment with the rest of our scientific worldview, especially our ability to make predictions about people's behaviour. I am less moved by this last point, since our social-scientific predictions are hardly ever of a deterministic kind; see my discussion in the next section. Saying that it is likely that an agent chooses a particular option – e.g., because we have evidence that it is the agent's most preferred option – is not the same as saying that it is impossible for the agent to do otherwise. The very fact that our standard descriptive theories of agency recognize certain forms of *akrasia* or even counterpreferential choice as a possibility shows that they take an agent's beliefs and preferences to underdetermine the agent's eventual choice. In any case, Gilboa and I are in agreement that rational choice theory is committed to representing several possible worlds, corresponding to different alternative actions, as being, in some sense, open to an agent.

if new scientific discoveries gave us reason to accept Premise 2 above – that determinism implies that (at any time) only one alternative course of action is *possible for the agent* – and not merely Premise 2* – that determinism implies that (at any time) only one future sequence of events is *physically possible*. The "wedge" that I have pushed between Premise 2* and Premise 2 would then no longer help us answer the incompatibilist challenge.

There are at least two conceivable developments that might lead us to accept Premise 2. One would be a successful reduction of psychology to physics – perhaps the dream of some neuroscientists – such that determinism at the physical level would straightforwardly imply determinism at the psychological level. Although neuroscientists have begun to identify a number of bridge laws that connect some specific cognitive phenomena with certain underlying patterns of brain activity, it is fair to say that a global reduction of psychology to physics is not in sight at this point.

The second conceivable development that might lead us to accept Premise 2 would be a paradigm shift towards a deterministic theory of psychology, even in the absence of a reduction of psychology to physics. However, despite frequent reports that certain behaviours can be explained by people's genes, their socio-economic backgrounds, upbringing, education, peer groups, and so on, the research findings in question are rarely able to attribute more than a certain *part* of the observed behavioural variance across people to those explanatory variables. These variables are shown at most to affect the *probabilities* of certain behaviours, and there is little reason to think that we are likely to arrive at a truly deterministic theory of psychology. Nonetheless, it is important to recognize that determinism at the psychological level would undermine free will in the sense discussed in this paper.

In fact, our practices of attributing responsibility to one another, both in informal social interactions and in legal contexts, have long adhered to the idea that if someone's actions can be shown to have been truly psychologically pre-determined, in the sense that only one action was possible for the agent, then this person's capacity for responsibility is seriously diminished. So determinism at the psychological level is indeed seen as a threat to free will in a way in which determinism at the level of fundamental physics is not. My argument explains why this intuition is correct.

7. Concluding remarks

Not everyone will agree with my argument, and some readers will find it misguided. They will say: "you may have identified a technical sense in which, relative to the epistemic limitations of the special sciences, free will can be 'defined into existence', but this hardly shows that free will is *truly* real". My response is that the special-science perspective is the only perspective from which we are likely to be able to defend free will, but that it is also a perspective from which we can reasonably be said to *have* free will. Few people doubt the reality of mental states such as beliefs, desires and intentions despite the fact that they are not part of the ontology of physics. These phenomena are only, and irreducibly, identifiable at the higher level, at which the special sciences operate. But their higher-level status does not undermine their reality. My claim is that free will is a higher-level phenomenon akin to those other phenomena and its higher-level status undermines its reality no less than it does in the case of those other phenomena.

I have argued that this position is supported by a naturalistic ontological attitude, non-reductive physicalism and our current best theories of agency, and so critics of my argument will have to explain why we should either deny ontological naturalism, or accept a reductive form of physicalism, or significantly revise our understanding of agency.[35] To give a "possibility proof" of how determinism at the physical level can co-exist with freedom at the agential one, I have sketched a toy model of the relationship between the lower and the higher level of a multi-level system and offered a simple semantic analysis of modal statements at both levels. The model should be of interest in its own right, independently of the issue of free will. It can also be used, for example, to illustrate the emergence of higher-level stochasticity – "chance" or "randomness" – in a deterministic world, provided the supervenience mapping from the lower to the higher level is sufficiently complex. (Probabilities can then be assigned to higher-level world histories – and the induced algebra of events – in terms of a density function defined on

[35] The naturalistic view that the question of whether we have free will is ultimately a scientific question about the nature of human decision-making processes is also held by Mark Balaguer in his recent book, *Free Will as an Open Scientific Problem* (Cambridge/MA: MIT Press, 2009), though the route by which he arrives at this conclusion is very different from the one taken in this paper. Balaguer defends a variant of libertarianism, which requires that certain neural events in the causal history of a human decision be suitably causally undetermined. Taking an explicitly anti-metaphysical view, Balaguer argues that much of the traditional compatibilism debate is irrelevant to the key questions about the nature of human freedom.

the subvenient lower-level histories.) This echoes the way in which statistical mechanics accounts for the emergence of stochasticity in a deterministic Newtonian world.[36] The special feature of the model introduced in this paper is that it combines ideas similar to those in statistical mechanics with ideas from modal logic.

However counterintuitive the present compatibilist view may seem at first sight, I believe that, on closer inspection, it is not as far-fetched as one might think. Imagine an article in next week's issue of *Nature* or *Science* that reports a big breakthrough in fundamental physics, including the finding that the universe is deterministic. How should we react to such news? Should we conclude that it challenges our understanding of the human condition as profoundly as the discovery of evolutionary theory did in the 19th century? Should we stop deliberating about what to do, and holding one another accountable for our actions? Should we release all murderers from prison? Or should we just go on with business as usual, thinking that the new discovery is an interesting quirk from physics?

Whatever we decide, it seems that giving up our conventional understanding of free will and revising the very fabric of how human society works would be an overreaction. The approach to free will offered in this paper shows why such a reaction is not warranted. The mildest revision of our technical vocabulary – namely the shift from physical to agential possibility in the analysis of free will – is sufficient to rehabilitate practically everything we conventionally believe and say about free will, even against the background of determinism. For this reason, my proposal seems to have common sense on its side.

In conclusion, I suggest that the best way to defend the compatibility of free will and determinism is to recognize that free will is not a physical phenomenon, but a higher-level phenomenon on a par with other familiar higher-level phenomena such as beliefs, desires and intentions. If we are searching for free will at the level of fundamental physics, we are simply searching in the wrong place.

[36] See, e.g., Jan Von Plato, "Probability and Determinism", *Philosophy of Science* 49, pp. 51-66 (1982), who argues that "it is possible to justify an objective interpretation of probabilities in a theory having as a basis the paradigmatically deterministic theory of classical mechanics" (p. 51).