Forthcoming in H. Beebee, C. Hitchcock, and H. Price (eds.), Making a Difference, Oxford (Oxford University Press).

My brain made me do it: The exclusion argument against free will, and what's wrong with it[1]

Christian List and Peter Menzies

December 2013, final version October 2014

> 'Did I consciously choose coffee over tea? No. The choice was made for me by events in my brain that I, as the conscious witness of my thoughts and actions, could not inspect or influence… The intention to do one thing and not another does not originate in consciousness – rather, it *appears* in consciousness, as does any thought or impulse that might oppose it.' (Harris 2012, pp. 7-8)

1. Introduction

There are at least two challenges that a scientifically oriented worldview may seem to pose for the idea that human beings have free will. The first is the familiar *challenge from determinism*. Suppose we accept

> **The alternative-possibilities thesis:** Someone's action is free *only if* he or she could have acted otherwise; and
>
> **The purported implication of determinism:** Determinism rules out the ability to act otherwise.

Then we must accept

> **The classical incompatibilist conclusion:** There can be no free actions in a deterministic world.

From a scientific perspective, determinism is still a live option, since a future theory of physics might represent the world as deterministic, even if current quantum physics seems to go against this.[2] This challenge has been extensively discussed and has generated a vast philosophical literature.[3]

However, there is a second challenge for free will, which is less discussed in philosophy, though, as we shall see, it resonates with recent discussions of

[1] We are very grateful for detailed written comments from Helen Beebee. We also thank Eddy Nahmias for helpful suggestions. Christian List's work was supported by a Leverhulme Major Research Fellowship. The slogan 'My brain made me do it' has been used as a title before. See, for example, Bloom (2006), Sternberg (2010), Mackintosh (2011), and Szalavitz (2012). This illustrates the attention that exclusion arguments against free will have received, albeit typically formulated in neuroscientific rather than philosophical terms. The slogan further appears in the title of a recent study of people's intuitions about free will, published after we finished this paper (Nahmias et al. 2014).

[2] At least, this is so on standard interpretations.

[3] For a recent contribution and further references, see List (2014).

neuroscience. This is what we may call the *challenge from physicalism*. It targets a slightly different thesis about free will:[4]

> **The causal-source thesis:** Someone's action is free *only if* it is caused by the agent, particularly by the agent's mental states, as distinct from the physical states of the agent's brain and body.

Suppose we conjoin this with

> **The purported implication of physicalism:** Physicalism rules out any agential or mental causation, as distinct from causation by physical states of the agent's brain and body.

Then we must accept

> **The source-incompatibilist conclusion:** There can be no free actions in a physicalist world.

Why might one think that physicalism rules out agential or mental causation – the purported implication of physicalism? The argument is a version of Jaegwon Kim's (1998) famous *exclusion argument*. The present version goes roughly as follows. Consider any action that is supposedly caused by an agent's mental states. Physicalism implies that those mental states supervene on (are determined by) physical states, most plausibly the agent's brain states. But then it is arguable that the real cause of the agent's action lies, not in the supervenient mental states, but in the underlying physical states.[5] The supervenient mental states are at most an epiphenomenon of the real, physical cause. If so, the action is not caused by the agent's mental states, contrary to our supposition, and thus, according to the causal-source thesis, does not qualify as free.[6]

In this paper, we critically assess this *exclusion argument against free will*. While the exclusion argument has received much attention in the literature on mental causation, it is seldom discussed in relation to free will, or it is mentioned only in passing.[7]

---

[4] Frankfurt-style examples are often taken to de-emphasize the alternative-possibilities thesis about free will and to strengthen our intuition that a key criterion for a free action is a causal source in the agent.

[5] This follows, in particular, from two widely accepted principles: the *causal closure of the physical world* and the *exclusion principle*, as explained in more detail below.

[6] To be precise, the action is not caused by the mental states *unless these are identical to the underlying physical states*. But any such mind-brain identity would go against the spirit of the causal-source thesis about free will and is something that all but the most reductively minded physicalists deny.

[7] Philosophers who have discussed the coherence of compatibilism with a physicalist worldview include Cover and O'Leary-Hawthorne (1996), O'Connor (2000), Merricks (2001), Roskies (2012), and Nahmias (2014). The three earlier works in this list focus on incompatibility arguments different from the one we formulate. One of these incompatibility arguments runs as follows: every action an agent performs supervenes on what the atoms composing his body do; no agent has a choice about what the atoms in his body do; therefore no agent has a choice about what actions he or she performs, in which case there is no free will. One problem with arguments of this kind is that they employ van Inwagen's notorious Beta Rule: if p entails q and no one has a choice about p, then no one has a choice about q. The arguments we formulate do not employ this rule. The more recent works we have cited are closer to ours. Roskies (2012) offers a naturalistic defence of free will against 'source-incompatibilist'

However, the argument expresses an idea that underlies the popular view that neuroscience, with its mechanistic picture of how the brain generates thought and behaviour, raises a serious challenge for free will. If our brains, rather than our conscious minds, cause our actions, how can those actions be free? This skeptical view is conveyed by the slogan 'my brain made me do it', suggesting that 'I' am not responsible. It is illustrated by our opening quote from the neuroscientist Sam Harris, who says that his choice of coffee over tea was 'made for [him] by events in [his] brain'; he was only a 'witness'. An analysis of the exclusion argument can help us assess this neuroscientific skepticism.[8]

We proceed as follows. In Section 2, we introduce two distinct versions of the exclusion argument against free will. In Section 3, we discuss several responses to it and suggest that most of them are not compelling. In Section 4, we explain our preferred response, which involves showing that a key premise – the *exclusion principle* – is false under what we take to be the most natural account of causation in the context of agency, namely the *difference-making account*. It says, roughly, that to be the cause of an effect is to be the difference-maker of that effect. We argue that, if we understand agential or mental causation in this way, we can uphold both the causal-source thesis and physicalism, while avoiding the source-incompatibilist conclusion. In developing this response, we draw on our earlier work on mental causation in List and Menzies (2009).[9] In Section 5, finally, we return to the topic of neuroscientific skepticism about free will.

2. The argument

We present two versions of the exclusion argument against free will. The first and simpler version explicitly invokes a physical causal-closure principle as a premise, while the second, more involved version replaces this with the premise that causation implies causal sufficiency. It is needless to say that we do not endorse all of the argument's premises. Indeed, our goal will be to explore which of them to give up.

2.1 The first version

Both versions of the argument have four premises. Here we begin with the first version:

---

challenges and discusses some debates about mental causation in relation to free will, arriving at a picture that seems broadly compatible with the one we defend (see also List and Menzies 2009 and List 2014). Nahmias (2014) offers a survey of several scientific challenges to free will and also arrives at compatibilist conclusions consistent with ours. Finally, in an unpublished manuscript, Wilson and Bernstein (2012) discuss some structural parallels between non-reductive physicalism as a response to the problem of mental causation and compatibilism as a response to the problem of free will. Their focus, however, is not so much on formulating and assessing an exclusion argument against free will or on exploring the non-reductive physicalist stance on free will, but rather on identifying parallels between the debate on mental causation and the debate on free will.

[8] Neuroscientific skepticism has increasingly been debated in the popular press. For discussion of these debates, see Roskies (2006) and, again, Nahmias (2014).

[9] For another, independent discussion of how to respond to Kim's original exclusion argument on the basis of a difference-making account of causation, see Raatikainen (2010).

Premise 1: An agent's action is free only if it is caused (in a relevant sense of causation *simpliciter*) by the agent's mental states.[10]

Premise 2: Any effect that has a cause has a sufficient physical cause (i.e., a causally sufficient physical condition) occurring at the same time.[11]

Premise 3: The agent's mental states are not identical to any physical states, but supervene on underlying physical states.

Premise 4: If an effect has a sufficient physical cause C, it does not have any cause C* (*simpliciter*) distinct from C occurring at the same time (except in cases of overdetermination).[12]

Premise 1 is a version of the causal-source thesis about free will, introduced above. Premise 2 is a very weak physical *causal-closure* principle, namely a conditional one: it asserts only that *if* an effect has a cause at a particular time, *then* it has a sufficient physical cause at that time. Note that the antecedent of this conditional refers to a cause *simpliciter*, while the consequent refers to a causally sufficient condition; these two notions need not coincide. Later we discuss the notions of *cause* and *sufficient cause* in more detail. Premise 3 is the central claim of non-reductive physicalism, according to which the relationship between mind and body, or the physical world more generally, is one of supervenience without identity: mental states supervene on underlying physical states, but are not reducible to them. Premise 4, finally, is a version of Jaegwon Kim's *exclusion principle*. It rules out the existence of two or more competing causes for the same event, but does not apply to cases of genuine causal overdetermination. Genuine overdetermination involves distinct causes that are unconnected or at most contingently connected, such as two assassins simultaneously shooting at the same target. The cases to which the exclusion principle applies are those in which the supposed rival causes are *necessarily* connected, especially via a supervenience relation. Examples are a brain state and a mental state that supervenes

[10] We discuss different notions of causation in Section 4 below.

[11] We use the term 'sufficient cause' as a shorthand for 'causally sufficient condition'. For the purposes of the present argument, we do not make any assumptions about whether a sufficient cause (i.e., a causally sufficient condition) automatically qualifies as a cause *simpliciter*. For example, we might say that, if the world is deterministic, the event of the big bang was causally sufficient for all subsequent events, and yet not prejudge the question of whether it should also count as a cause *simpliciter* of all subsequent events. Indeed, on the account of causation that we ultimately defend (the difference-making account), causally sufficient conditions are conceptually distinct from causes. As discussed further below, Premise 2 does not by itself imply or presuppose determinism. Consistently with Premise 2, there could be events without any causes in an indeterministic world. We also discuss a version of our argument that invokes a probabilistic notion of causation.

[12] To avoid certain trivial counterexamples, one might fine-tune this premise by referring to a *minimal* sufficient physical cause in its antecedent. There can easily be distinct sufficient physical causes for the same effect, occurring at the same time, and so it can happen that C is a minimal sufficient physical cause for E and C* is a non-minimal one which entails C. Even if we stipulated that only C but not C* qualifies as a cause *simpliciter* for E, Premise 4 would be violated in its original form, since it would imply that C*'s being a sufficient physical cause for E excludes C, which is distinct from C*, from being a cause *simpliciter*. With the minimality restriction in the antecedent, Premise 4 would be satisfied. For simplicity, however, we use the original, unrestricted formulation in the main text. Later we discuss non-trivial counterexamples to Premise 4.

on it, which might both be candidate causes of the same effect, such as an agent's action. Note that, in its present formulation, the exclusion principle refers to a sufficient cause in its antecedent clause, and to a cause *simpliciter* in its consequent clause.[13]

Premises 1 to 4 entail

> **The conclusion:** There are no free actions.

To see this, suppose a particular action is free. By Premise 1, it is caused by the agent's mental states; call the relevant set of mental states C*. By Premise 2, since the action has a cause (namely C*), it has a sufficient physical cause occurring at the same time; call it C. By Premise 3, C* is not identical to C, but supervenient on C; we assume that C is specified sufficiently richly to include the supervenience base of C*. In light of the supervenience relationship between C and C*, we are not dealing with a case of causal overdetermination. Hence, by Premise 4, C's being a sufficient cause for the action excludes C* from being a cause, a contradiction.

2.2 The second version

The second version of the argument replaces Premise 2 with

> **Premise 2*:** Causation implies causal sufficiency.[14]

Again, the resulting premises entail

> **The conclusion:** There are no free actions.

To see this, we first require a preliminary result:

> **Lemma:** If C* is causally sufficient for some effect E, and C* supervenes on C, then C is causally sufficient for E.

The proof of this lemma is straightforward. Suppose C* is causally sufficient for E, and C* supervenes on C. Assume, for a contradiction, that C is not causally sufficient for E. Then C could occur without E. But the occurrence of C necessitates the occurrence of C*, which is sufficient for E, a contradiction.

To show that Premises 1, 2*, 3 and 4 entail that there are no free actions, suppose a particular action is free. By Premise 1, it is caused by the agent's mental states; call the relevant set of mental states C*. By Premise 2*, since C* is a cause of the action,

---

[13] Sufficient causes need not compete with one another. As we have already noted, it is entirely possible for C to be causally sufficient for E, and for C*, which is distinct from C but entails C, to be causally sufficient for E as well.

[14] Whether Premise 2* is plausible depends on how causation and causal sufficiency are understood. For the premise to be plausible, both causation and causal sufficiency might have to be understood as referring to the relevant background circumstances. Below we briefly assess different interpretations of Premise 2*.

it is a sufficient cause. By Premise 3, the agent's mental states C* are not identical to any physical states, but supervenient on underlying physical states; call the relevant set of physical states C. By our lemma, since C* is causally sufficient for the agent's action, and C* supervenes on C, C is itself causally sufficient for the agent's action. In light of the supervenience relationship between C and C*, we are not dealing with a case of causal overdetermination. Hence, by Premise 4, C's being a sufficient cause for the action excludes C* from being a cause, a contradiction.

3. Some responses to the argument

We can avoid the conclusion that there are no free actions only by giving up at least one of the premises of each version of the argument. Let us briefly discuss the options, considering each premise on its own terms. Although we are ultimately interested in responses to the argument that are consistent with a scientifically oriented worldview, we note that some of the premises do not, by themselves, imply or presuppose any version of physicalism. In this sense, the premises are quite ecumenical.

3.1 Giving up Premise 1

Recall that Premise 1 says that a necessary condition for an action to be free is that it is caused by the agent's mental states. This is a version of the causal-source thesis and captures an important intuition about free will. It is even weaker than the causal-source thesis as formulated in the introduction. While that thesis included the requirement that the mental states causing the agent's action be distinct from the physical states of the agent's brain and body, Premise 1 does not include this requirement. Thus, Premise 1 is consistent with the possibility that the relevant mental states could be identical to underlying brain and bodily states. So, even proponents of reductive physicalism should have little grounds for rejecting Premise 1 by itself.

Furthermore, just as Premise 1 does not rule out an *identity* between mind and brain, so it does not rule out a complete *disconnect* between the two. Without additional assumptions, it allows the mental states that cause the agent's action to be non-supervenient on any physical states. Thus Premise 1 on its own is compatible even with interactionist dualism of the traditional Cartesian sort.

The only way to reject this premise while holding on to the idea that free actions have a causal source in the agent would be to insist on a form of *agent causation* under which actions are caused not by the agent's mental states but by some other aspect of the agent. This idea, however, seems metaphysically mysterious, and we set it aside. In sum, we think Premise 1 is hard to give up.

*3.2 Giving up Premise 2 or 2**

We have considered two versions of the second premise, 2 and 2*. As already noted, Premise 2 is a very weak physical causal-closure principle. It only asserts the existence of sufficient physical causes for those events that have a cause. This is

entirely consistent with the occurrence of *non-physical* events that have no cause and even with the occurrence of *physical* events that have no cause (e.g., genuinely uncaused events in an indeterministic world). Physicalists, whether of a reductive or non-reductive kind, should have no problem with this premise. We would only have to relax Premise 2 if we accepted physical indeterminism *and* were nonetheless prepared to say that some physically un-determined events have causes. We turn to the issue of indeterministic causation in our discussion of Premise 2* below.

Non-physicalists may find it harder to accept Premise 2. Interactionist dualists would presumably reject it, though their metaphysical picture does not fit with a scientific worldview, so we set it aside. Naturalistically minded dualists might accept Premise 2 against the background of a nomological supervenience relation between the physical and the non-physical, and thus accept that, *given the laws of our world*, any event that has a cause has a sufficient physical cause, understood as a physical condition whose presence would *nomologically* necessitate the event in question. In sum, Premise 2 seems hard to give up unless we are prepared to depart significantly from a scientifically oriented worldview.

Premise 2*, unlike Premise 2, requires no form of physical causal closure and instead constrains the notion of causation. It says that any event C that is the cause of an effect E (in the sense of causation *simpliciter*) is also causally sufficient for E. Whether this principle is acceptable or not depends on how we understand the notions of *causation* and *causal sufficiency*. If causal sufficiency merely required that if C were to occur, then E would occur (a *counterfactual conditional*), then Premise 2* would be hard to deny under almost any notion of deterministic causation. By contrast, if causal sufficiency required nomological necessitation (something akin to the *strict conditional*: necessarily, if C then E) while causation required counterfactual difference-making, Premise 2* would be questionable. Difference-making causes make true two counterfactual conditionals: first, if C were to occur, then E would occur; and second, if C were not to occur, then E would not occur. But they do not generally make true the strict conditional: necessarily, if C then E.

Finally, it is possible to formulate the second version of our exclusion argument in terms of a probabilistic notion of causation: any reference to causal sufficiency in the argument must then be replaced by a reference to high conditional probability. Suppose, in particular, we replace the premise that causation (*simpliciter*) implies causal sufficiency with the premise that causation (*simpliciter*) implies high conditional probability: i.e., C's causation of E implies that Pr(E|C) is high. Then it is easy to demonstrate that high conditional probability is transmitted across supervenience when a plausible 'Markov condition' obtains: if (i) Pr(E|C*) is high, (ii) a subvenient event C entails C*, and (iii) C* screens off C from E (i.e., Pr(E|C*&C) = Pr(E|C*)), then Pr(E|C) = Pr(E|C*). Given this new premise and the new lemma, the causal source argument proceeds much as above.

In sum, although Premise 2* can be denied under some assumptions about causation and causal sufficiency, it would seem unsatisfactory to let the vindication of free will hinge on those assumptions.

3.3 Giving up Premise 3

Giving up Premise 3 would be to deny non-reductive physicalism. Kim advocates this, espousing reductive physicalism, which involves the denial of non-identity. But we think that this would be an unsound response to the exclusion argument against free will, for at least two reasons.

First, the multiple-realizability objection to reductive physicalism seems broadly correct. For all we know, mental states are more coarse-grained than their subvenient brain states and can be physically realized in multiple ways. Hence mental states cannot be identified with their physical realizers.

Second, even if we set the issue of multiple realizability aside, most proponents of the causal-source thesis about free will are likely to accept a non-reductive view about the relationship between brain and mind. This is because they wish to emphasize that free actions are those caused by an agent's mental states, *qua* rationalizing intentional states, not *qua* physical states of the brain. Free will, on this picture, is a higher-level, agential phenomenon, not a lower-level, physical one.[15] It is plausible to think that free will presupposes rational capacities for controlling one's actions. While one can rationally control one's conscious intentional states, one cannot rationally control one's brain states, in part because one is seldom aware of them. So, free action seems to require causal explanation at the intentional level rather than at the physical level. This, in turn, makes the non-identity part of Premise 3 hard to give up.

Another way to relax Premise 3 would be to maintain that mental states need not supervene on physical states. Consistently with the other premises, we could then conclude that free actions are genuinely causally overdetermined: they have both an agential cause and a physical cause, which stand at most in a contingent relationship to one another. (The lack of a necessary relationship is important, since in the presence of a necessary relationship we would not be able to apply the exemption clause in Premise 4, which permits cases of overdetermination.)

Giving up supervenience of the mental on the physical, however, seems not very satisfactory. Not only would the present route involve a significant departure from a scientifically oriented worldview, but the exclusion argument against free will would be reinstated if we strengthened Premise 4 by dropping its exemption clause (referring to causal overdetermination). In sum, we see little promise in dropping Premise 3.

---

[15] On free will as a higher-level phenomenon, see also List (2014). The present picture is further supported by the discussion in Roskies (2012), who emphasizes the role of psychological, as opposed to physical, control variables for intentional action. By a control variable, Roskies means roughly a variable which, when changed by certain interventions, leads to systematic changes in other variables. For discussion of the notion, see Hitchcock and Woodward (2003).

3.4 Giving up Premise 4

As should be evident by now, if we wish to resist the exclusion argument against free will without sacrificing other central tenets of a scientifically oriented worldview, we must give up Premise 4: the exclusion principle. Drawing on our previous work on mental causation, we now show that this principle is false if we accept the account of causation that is arguably most natural in the context of agency: the difference-making account.

4. Our diagnosis of the argument's flaw

4.1 Causation

There are at least two fundamentally different ways in which the notion of causation can be understood: as 'production' or as 'difference-making'.[16] Each of these labels corresponds to a family of accounts of causation.

On a production account, to be the cause of an effect is to be the producer of that effect, in some metaphysical sense of production. Causation here involves a causal 'oomph', i.e., the production of an outcome through some causal force or power, on the model of a billiard ball's causing the motion of another by transmitting a force on impact.

On a difference-making account, by contrast, causation is a form of counterfactual or probabilistic dependence: to be the cause of an effect is to be the difference-maker of that effect. For convenience, we here spell this out in counterfactual terms: C causes E if and only if two conditionals are satisfied, as already mentioned above:

> **The positive conditional:** If C were to occur, then E would occur.
>
> **The negative conditional:** If C were not to occur, then E would not occur.

Arguably, a difference-making account is more in line with the practices of causal attribution and explanation in the sciences than a production account is. When scientists obtain evidence for counterfactual difference-making on the basis of careful experimental or statistical controls, they usually interpret this as evidence for causation. On a production account, which involves the idea of a 'causal oomph', such an interpretation would involve a leap of faith: evidence for counterfactual difference-making is not automatically evidence for a causal 'oomph'. On a difference-making account, on the other hand, evidence for counterfactual difference-making is naturally evidence for causation, because causation is just counterfactual difference-making.

---

[16] See, e.g., Armstrong (2004), Hall (2004), and Kim (2005). Hall speaks of 'production' and 'dependence'.

A difference-making account also fits well with the connection that is frequently made between causation and intervention in a system.[17] For C to be the cause of E, on this picture, it must be the case that interventions on C make a difference to E. Further, as we have argued elsewhere (List and Menzies 2009), the most natural way to spell out the idea of mental causation is to say that an agent causes an action if and only if his or her mental state is the difference-maker of the action.

Conceptually, producing causes and difference-making causes must be distinguished from one another. When a flask of boiling water breaks due to the pressure, the producing cause may well be the motion of a specific subset of the water molecules; yet, the difference-making cause is the boiling of the water. Only the latter but not the former, satisfies the two (positive and negative) conditionals stated above.[18] We return to this example below.

Causally sufficient conditions – whether or not they qualify as producing causes – are not the same as difference-making causes. A man's taking a contraceptive pill is causally sufficient – in some vacuous sense – for his not becoming pregnant, but there is no genuine causal relation here: neither of the producing kind, nor of the difference-making kind.[19]

In sum, we suggest that the best interpretation of causation in the context of agency is a difference-making one: to say that an action is caused by the agent's mental states is to say that those mental states are the difference-making causes of the action.[20] What does this imply for the exclusion argument against free will?

4.2 The falsity of the exclusion principle

Recall that the exclusion principle states that if an effect has a sufficient physical cause C, it does not have any cause C* distinct from C occurring at the same time, except in cases of overdetermination. Kim defended this principle on the basis of a production account of causation. The idea is that, except in cases of genuine overdetermination, the causal responsibility for any effect must be uniquely attributable: the same effect cannot be due to two or more simultaneous but competing sources of causal power. Regardless of whether this principle is plausible under an account of causation as production, it is easy to see that it is false when causation is understood as difference-making.

---

[17] See Pearl (2000) and Woodward (2003).

[18] In this particular example, we are broadly in agreement with Jackson and Pettit's analysis (1990). See also Pettit (2013).

[19] Note that, since a man (under standard assumptions) can never become pregnant, his taking a contraceptive pill cannot change that fact and hence will qualify as a sufficient cause for his not becoming pregnant, no matter whether we interpret causal sufficiency in nomological terms, counterfactual terms, or probabilistic terms. Further, note that the claim that causal sufficiency does not imply causation is consistent with the reverse claim that causation implies causal sufficiency, asserted by Premise 2*, though we need not commit ourselves to the latter claim either.

[20] We set aside cases of preemption and overdetermination, for which the simple analysis of difference-making causation in terms of the positive and negative conditionals does not work. A more sophisticated analysis is required for such cases.

Again, consider the flask of boiling water. Its full molecular microstate at the time of the breaking may well be a sufficient cause for the breaking, and the boiling of the water supervenes on that microstate. Yet, it is the boiling that is the difference-making cause of the breaking, not the underlying microstate. If the boiling had occurred, but had been realized by a slightly different microstate, the flask would still have broken, and if the boiling had not occurred, the flask would have remained intact. So, the positive and negative conditionals for difference-making are satisfied when C is the boiling of the water and E is the breaking of the flask. By contrast, although it is true that if the microstate of the flask had been exactly as it was, the flask would have broken, it is not true that if the microstate had been slightly different, the flask would have remained intact. The boiling could have been realized in many different ways, through different configurations of molecular motion, and would still have led the flask to break. So, while the *positive* conditional for difference-making is satisfied when C is the microstate and E is the breaking of the flask, the *negative* conditional is not. This shows that the *difference-making* cause for the breaking of the flask is the boiling event, not the microstate on which it supervenes. Nonetheless, the microstate of the flask is *causally sufficient* for the breaking. Consequently, we have a counterexample to the exclusion principle.

Similarly, we have argued elsewhere (List and Menzies 2009) that when an agent intentionally moves his or her arm, the difference-making cause of the action is (plausibly) not the subvenient brain state, but the supervenient intention (for a similar argument, see also Raatikainen 2010). Only the intention, but not the brain state, satisfies the two conditionals for difference-making. If the intention were present, the action would be performed, and if the intention were absent, it would not. The brain state, by contrast, satisfies only the positive conditional, but arguably not the negative one. If the precise realizing brain state were absent, but the same intention were realized by another brain state, the action would presumably still be performed. Accordingly, the exclusion principle is violated here: the brain state is causally sufficient for the action, and the intention supervenes on it; yet, on the difference-making account, it is the intention, not the brain state, that is the cause of the action. These considerations show that the exclusion principle is false when causation is understood as difference-making.

In sum, we reject Premise 4 and can therefore consistently accept Premises 1 to 3 while still holding the view that there can be free actions. This, we believe, is the most compelling response to the exclusion argument against free will.

5. Neuroscientific skepticism about free will revisited

Neuroscientific skepticism about free will is the view that advances in neuroscience, especially discoveries of the neural causes of thought and behaviour, raise serious challenges for free will. For any of our purportedly intentional actions, we seem increasingly warranted in saying: ‘My brain made me do it. Hence it was not my own free choice.’ Thus Harris (2012) concludes:

> 'Free will *is* an illusion. Our wills are simply not of our own making. Thoughts and intentions emerge from background causes of which we are unaware and over which we exert no conscious control. We do not have the freedom we think we have.' (p. 5)

Similar claims can be found in the popular writings of other neuroscientists, who, like Harris, attempt to startle their readers with the claim that free will is an illusion and that human action is a consequence of physical processes beyond our control. Gazzaniga (2011) is another prominent example of this genre. To assess such neuroskeptical claims, we need to make the argument for them more precise.[21]

5.1 The neuroskeptical argument

The argument that proponents of neuroskepticism tend to invoke – albeit often implicitly – has two premises:

> **The purported exclusion of free will by neural causes:** If an agent's choices and actions are wholly caused by neural states and processes that are inaccessible to his or her consciousness, then these choices and actions are not free.
>
> **The thesis of neural causation:** Human choices and actions are wholly caused by neural states and processes that are inaccessible to the agent's consciousness.

These premises then support

> **The neuroskeptical conclusion:** Human choices and actions are not free.

What should we say about this argument? The argument is certainly valid, and its premises seem at first sight plausible.

Consider the first premise, the purported exclusion of free will by neural causes. To the extent that the causes of an agent's choices and actions bypass the conscious mental states that are supposed to play a role in deliberation, it would appear that these choices and actions are indeed not free. As Nahmias (2006) has shown through survey evidence, this thesis is something that most ordinary people who are not trained in philosophy believe. Thus the first premise is a widely accepted assumption about free will.

Similarly, the second premise, the thesis of neural causation, is plausible, as it appears to be supported by a growing body of experimental work. Libet's classic study of the neuronal readiness potentials that precede conscious intentions to perform actions is a widely cited piece of evidence. Libet (1983) showed that the neuronal activity leading to the performance of an action tends to begin several hundred milliseconds before a

[21] For a helpful review of scientific challenges to free will, see Nahmias (2010).

subject appears to be consciously aware of his or her intention to act. Although the precise interpretation of this finding remains controversial, Libet's experiment has been replicated by others, in a number of variations. In a particularly dramatic study of the neural correlates of intention formation, Haynes and colleagues (2007) were able to use brain-scan data to predict subjects' choices between two actions 7-10 seconds before the action took place.[22] All these findings seem consistent with the claim that human choices and actions are ultimately the result of subconscious processes beyond an agent's control.

Despite their initial plausibility, however, the two premises do not withstand closer scrutiny. Let us discuss them in turn.

5.2 The purported exclusion of free will by neural causes

The first premise – the purported exclusion of free will by neural causes – is most plausible when it is interpreted as relying on an exclusion argument of the kind we have investigated in previous sections. To show that the premise holds, one might argue as follows. Let us assume that an agent's actions are wholly caused by neural states and processes; it then follows from the non-identity of mental states and neural states, together with an exclusion principle, that these actions are not caused by any mental states occurring at the same time; and so, by the causal-source thesis, the actions are not free.

But our preceding discussion should alert us to the fact that this reasoning depends on the relevant exclusion principle. If this principle states that the existence of a physical *sufficient* cause for an action excludes any simultaneous mental state from being a *difference-making* cause of that action – as formulated in Premise 4 – we have good reason to reject it. We showed in the last section that *sufficient* causes at the physical level can co-exist with distinct, higher-level *difference-making* causes of the same effects. For example, the fact that a specific molecular microstate of a gas is sufficient to break the walls of its container is consistent with the fact that the macrostate of the gas's pressure on the walls is the difference-making cause of the breaking.

In the case of human action, the same is true of sufficient causes at the neural level and difference-making causes at the mental level. As we have seen, the existence of a neural state that is causally sufficient for some action is consistent with the existence of a difference-making cause at the mental level, such as the agent's intention. Thus the reasoning in support of the first premise of the neuroskeptical argument does not go through.

Alternatively, suppose we try to support that premise by reinterpreting the exclusion principle as follows:

[22] For a detailed critical discussion to which we are indebted, see Nahmias (2014).

> **Reinterpreted exclusion principle:** If an effect has a *difference-making cause* C at the physical level, it does not have any other *difference-making cause* C* at the mental level, occurring at the same time.

This is distinct from the exclusion principle on which we have focused so far, which refers to a *sufficient* cause, not a difference-making cause, in the antecedent. Therefore our rejection of the earlier principle – in the form of Premise 4 – does not carry over to the present, reinterpreted version. The present principle only says that there cannot simultaneously exist more than one *difference-making* cause for the same effect, at different levels. This is consistent with the possibility that a lower-level *sufficient* cause might co-exist with a higher-level *difference-making* cause. Indeed, as we have shown in List and Menzies (2009), there are conditions under which the reinterpreted exclusion principle holds, despite the failure of the original principle.

But now it is reasonable to ask: Does the reinterpreted principle vindicate the neuroskeptical argument? We reply: Not automatically, because the skeptic still has to establish the second premise of the argument, the thesis of neural causation, interpreted in terms of difference-making causation. In other words, the skeptic has to show that there are difference-making *neural* causes for all actions. It is not enough to show this in a single instance. So, let us turn to the thesis of neural causation.

5.3 The thesis of neural causation

As already noted, when we understand causation as difference-making, we are likely to conclude that the cause of an agent's action is not the agent's brain state, but his or her mental state. Only the supervenient mental state, but not the subvenient brain state, may satisfy the two conditionals for difference-making. Recall, in particular, that the realizing brain state plausibly violates the negative conditional: if it were absent, the action might still be performed, provided the same mental state is realized by some other brain state.

More generally, we can identify conditions under which a supervenient event alone, rather than its physical realizer, is the difference-making cause of an effect. The following result holds (List and Menzies 2009):

> **Proposition:** A supervenient event C* is the difference-making cause of an effect E, and its subvenient realizer C is not, *if and only if* C* and E satisfy the positive and negative conditionals for difference-making *and* this causal relationship is realization-insensitive. (*Realization-insensitivity* means that the effect E continues to occur under some small perturbations in the physical realization of the cause C*; formally, E occurs in some closest not-C-worlds that are C* worlds.)

Figure 1 (adapted from List and Menzies 2009) provides an illustration. The figure shows a space of possible worlds. The small dot in the central circle corresponds to the actual world. The concentric circles around it correspond to increasingly distant possible worlds. Any worlds within the same circle (i.e., either within the inner-most

circle, or within the second circle but outside the inner-most circle, or within the third circle but outside the second, and so on) are deemed to be equidistant from the actual world. The large half-oval region on the left-hand side corresponds to the set of worlds in which the supervenient event C* occurs. The smaller half-oval with the diagonal lines corresponds to the set of worlds in which C* has the physical realizer C. The shaded region in the centre corresponds to the set of worlds in which the effect E occurs.

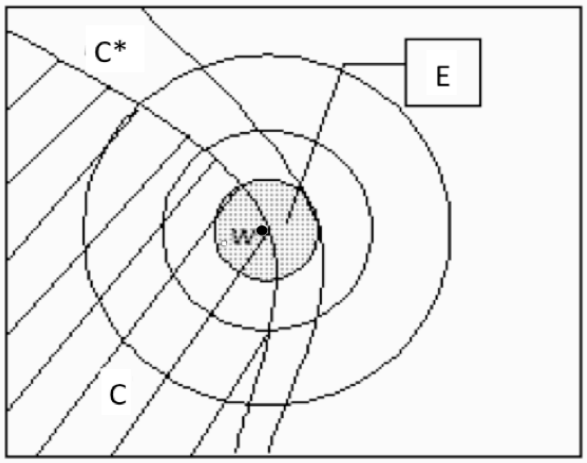

Figure 1: Realization-insensitive causation

It is easy to see that, under this configuration, C* but not C is a difference-making cause of E: relative to the actual world, E is present in all nearest possible worlds in which C* occurs, and absent in all nearest possible worlds in which C* does not occur. By contrast, while E is present in all nearest possible worlds in which C occurs, it is not absent in all nearest possible worlds in which C does not occur; indeed, E continues to occur in those worlds in the central circle in which C* has a different physical realizer. In sum, C*, but not C, is a difference-making cause of E, and the causal relationship is realization-insensitive.

Here, contrary to the familiar idea of a lower-level cause excluding a higher-level one ('upwards exclusion'), the supervenient event C* excludes the subvenient event C from being a difference-making cause of E: a 'downwards exclusion' result. The bottom line is that the skeptic cannot take the thesis of neural causation for granted; the existence of a mental difference-making cause of some action may well exclude the existence of any underlying physical difference-making cause of the same action.

These conditions for 'downwards exclusion' are fully consistent with the reinterpreted exclusion principle we have discussed. Of course, it is an empirical question whether the difference-making cause of an agent's action is some mental state or not. However, psychological experimentation should enable us to answer this question, by establishing whether it is true that if an agent were to possess a particular mental state, he or she would perform the action, and if the agent were not to possess that mental state, he or she would not perform it.

When there is such a causal relationship and this relationship is *realization-insensitive*, our 'downwards exclusion' result implies that no subvening neural cause can be a difference-making cause of the action. The control variables for an agent's actions may well be the relevant mental states rather than their physical realizers.[23]

5.4 Concluding remarks

Although the reinterpreted exclusion principle goes some way towards vindicating the first premise of the neuroskeptical argument, the possibility of 'downwards exclusion', which is consistent with the principle, goes an equal distance in limiting the credibility of the second premise. If mental states, rather than their physical realizers, are the difference-making causes of an agent's actions, we must reject the thesis of neural causation, the second premise of the neuroskeptical argument.

In sum, the neuroskeptical argument against free will is unsound. The argument depends crucially on the plausibility of an exclusion principle underlying its first premise. When this principle is formulated in terms of the compatibility of *sufficient* physical causes with *difference-making* mental causes, we have good reason to reject it. When the principle is formulated in terms of the compatibility of *difference-making* physical causes with *difference-making* mental causes, the principle has some credibility under suitable conditions. But the support it lends to the first premise of the argument is counterbalanced by the doubt that the possibility of 'downwards exclusion' casts on the second.[24]

References

Armstrong, D. (2004). 'Going through the Open Door: Counterfactual vs. Singularist Theories of Causation', in Collins et al., 445-457.

Bloom, P. (2006). 'My brain made me do it', *Journal of Cognition and Culture*, 6(1-2): 209-214.

Collins, J., N. Hall, and L. A. Paul (eds) (2004). *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Cover, J., and J. O'Leary-Hawthorne (1996). 'Haecceitism and Anti-Haecceitism in Leibniz's Philosophy', *Noûs*, 30(1): 1-30.

Gazzaniga, M. (2011). *Who's in Charge: Free Will and the Science of the Brain*. New York: Harper Collins.

Hall, N. (2004). 'Two Concepts of Causation', in Collins et al., 225-276.

[23] On this point, see also Roskies (2012) and footnote 15 above.

[24] For a recent empirical study suggesting that most people are not persuaded by a certain common form of neuroscientific skepticism about free will (which claims that the possibility of perfect prediction of behaviour based on neural information undermines free will), see Nahmias et al. (2014).

Haynes, J.-D., K. Sakai, G. Rees, S. Gilbert, C. Frith, and R. E. Passingham (2007). 'Reading Hidden Intentions in the Human Brain', *Current Biology*, 17: 323-328.

Harris, S. (2012). *Free Will*. New York: Simon & Schuster.

Hitchcock, C., and J. Woodward (2003). 'Explanatory Generalizations, Part II: Plumbing Explanatory Depth', *Nous*, 37(2): 181-199.

Jackson, F., and P. Pettit (1990). 'Program Explanation: A General Perspective', *Analysis*, 50: 107-117.

Kim, J. (1998). *Mind in a Physical World*. Cambridge, MA: MIT Press.

Kim, J. (2005). *Physicalism, Or Something Near Enough*. Princeton: Princeton University Press.

Libet, B. (1983). 'Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness Potential): the Unconscious Initiation of a Freely Voluntary Act', *Brain*, 106: 623-642.

List, C. (2014). 'Free Will, Determinism, and the Possibility of Doing Otherwise', *Nous*, 48(1): 156-178.

List, C., and Menzies, P. (2009). 'Non-Reductive Physicalism and the Limits of the Exclusion Principle', *Journal of Philosophy*, 105: 475-502.

Mackintosh, N. (2011). 'My brain made me do it', *New Scientist*, 212(2843): 26-27.

Merricks, T. (2001). *Objects and Persons.* New York: Oxford University Press.

Nahmias, E. (2006). 'Folk Fears about Freedom and Responsibility: Determinism vs. Reductionism', *Journal of Cognition and Culture*, 6: 215-37.

Nahmias, E. (2010). 'Scientific Challenges to Free Will', in T. O'Connor and C. Sandis (eds.), *A Companion to the Philosophy of Action*, Oxford: Wiley-Blackwell, 345-356.

Nahmias, E. (2014). 'Is Free Will an Illusion? Confronting Challenges from the Modern Mind Sciences', in W. Sinnott-Armstrong (ed.), *Moral Psychology, Vol. 4, Free Will and Moral Responsibility*, Cambridge/MA: MIT Press, 1-26.

Nahmias, E., Shepard, J., and Reuter, S. (2014). 'It's OK if "my brain made me do it": People's intuitions about free will and neuroscientific prediction', *Cognition*, 133(2): 502-516.

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

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge: Cambridge University Press.

Pettit, P. (2013). 'The Program Model, Difference-makers, and the Exclusion Problem', unpublished paper.

Raatikainen, P. (2010). 'Causation, Exclusion, and the Special Sciences', *Erkenntis*, 73(3): 349-363.

Roskies, A. (2006). 'Neuroscientific Challenges to Free Will and Responsibility', *Trends in Cognitive Science*, 10(9): 419-423.

Roskies, A. (2012). 'Don't Panic: Self-authorship without Obscure Metaphysics', *Philosophical Perspectives*, 26: 323-342.

Sternberg, E. J. (2010). *My Brain Made Me Do It: The Rise of Neuroscience and the Threat to Moral Responsibility*. Amherst, New York: Prometheus Books.

Szalavitz, M. (2012). 'My Brain Made Me Do It: Psychopaths and Free Will', *Time*, August 17, 2012.

Wilson, J., and S. Bernstein (2012). 'Free Will and Mental Causation', unpublished manuscript.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.