Revisiting McKay and Johnson's counterexample to (β)*

Pedro Merlussi

Centre for Logic, Epistemology and the History of Science - University of Campinas

November 4, 2021

Abstract

In debates concerning the consequence argument, it has long been claimed that McKay and Johnson (1996) demonstrated the invalidity of rule (β). Here, I argue that their result is not as robust as we might like to think. First, I argue that McKay and Johnson's counterexample is successful *if* one adopts a certain interpretation of "no choice about" *and* if one is willing to deny the conditional excluded middle principle. In order to make this point I demonstrate that (β) is valid on Stalnaker's theory of counterfactuals. This result is important and should not be neglected, I argue, because there is a particular line of objection to the revised formulations of the consequence argument that does not succeed against the original version.

*I would like to express gratitude to two anonymous reviewers from this journal who provided invaluable feedback on this paper. I am grateful to the reviewers from another journal who have made a number of criticisms to the original manuscript resulting in a much improved article. I am also grateful to Fabio Lampert, Julio de Rizzo, Nancy Cartwright, Matthew Tugby, Carolina Sartorio and Helen Beebee.

1 Introduction

Is determinism compatible with free will? If the consequence argument is sound, the answer is "no". van Inwagen's formulation (1983: 56) runs as follows:

> If determinism is true, then events about our actions are the consequence of a proposition L stating the laws of nature together with a proposition P_0 about the distant past.[1] But we have no choice about whether L is true. And we have no choice about whether P_0 is true. Hence, we have no choice about the consequences of those things, including our actions.

The key of the argument is the assumption that "no-choice about" transfers across conditionals: if P is true and no one has any choice about whether P is true, and $P \supset Q$ is true and no one has any choice about whether $P \supset Q$ is true, then Q is true and no one has any choice about whether Q is true. This is what the (in)famous rule (β) says in the following standard modal formulation. Let $\mathbf{N}P$ stand for "P and no one has or ever had any choice about whether P is true" (van Inwagen 1983: 93-95). Here is the modal version:

(α) $\Box\phi \vdash \mathbf{N}\phi$

(β) $\mathbf{N}\phi, \mathbf{N}(\phi \supset \psi) \vdash \mathbf{N}\psi$

1	$\Box((L \wedge P_0) \supset P)$	determinism
2	$\Box(L \supset (P_0 \supset P))$	modal logic, 1
3	$\mathbf{N}(L \supset (P_0 \supset P))$	α, 2
4	$\mathbf{N}L$	premise
5	$\mathbf{N}P_0$	premise
6	$\mathbf{N}(P_0 \supset P)$	β, 3, 4
7	$\mathbf{N}P$	β, 5, 6

[1] As we now know, P_0 needs to be a proposition about the *distant* past (Campbell 2007). I am grateful to an anonymous referee for drawing my attention to this.

Although there is still a great deal of debate over the consequence argument, the modal formulation above is too problematic. Orthodoxy tells us that (β) is *invalid*.

> [...] even though Beta seems intuitively valid, Beta turned out to be invalid. In a paper published in 1996, McKay and Johnson demonstrated the invalidity of Beta in two steps. (Vihvelin 2013: 160)

Rule (β) does seem valid because $\mathbf{N}$ is some sort of unavoidability operator, and one might expect it to be something pretty similar to the box of logical necessity. "No-choice about" should transfer across material conditionals just like the box does (if it's necessary that P and necessary that $P \supset Q$, then it's necessary that Q). However, McKay and Johnson are credited to have demonstrated that rule (β) is invalid (see Huemer 2000; van Inwagen 2000; Vihvelin 2017).[2] So, contrary to appearances, (β) is not like the box.

Yet my claim in this paper is that McKay and Johnson's result is not robust enough: What they demonstrated is that rule (β) is invalid *if* we make certain assumptions about counterfactuals *and* assume a certain interpretation of "no-choice about" (namely, the counterfactual sufficiency interpretation), as I shall present it in section 2. Moreover, I argue in section 3 that things are different if one adopts a Stalnakerian view of counterfactuals. Given the same interpretation of "no-choice about" where (β) fails (for example, on Lewis' view), I demonstrate that (β) is valid on Stalnaker's view. What is at stake here, I shall argue, is the conditional excluded middle.

In section 4 I discuss the relevance of this result in a more general context concerning the cogency of the consequence argument. Since (β) is taken to be invalid, most incompatibilists adopt a revised definition of "no choice about" that makes some (β)-like principle valid, which is so regardless of controversial

[2] Perhaps the only exception is Blum (2000). Even so, he argues for the paradoxical conclusion that "$\mathbf{N}$ ought to be, and yet ought to fail to be, agglomerative" (Blum 2000: 286).

assumptions about counterfactuals. However, while this move has its merits, we will see that it also makes the consequence argument more vulnerable to objections that did not affect the original version. I shall argue that there is a particularly threatening objection to the updated consequence arguments that leaves the original formulation unscathed. In other words, the fact that rule (β) is valid on Stalnaker's theory of counterfactuals should not be neglected.

2 McKay and Johnson's counterexample to (β)

McKay and Johnson's counterexample to (β) has indeed two steps. The first one is a counterexample to agglomeration:

(Agglomeration) $\mathbf{N}\phi, \mathbf{N}\psi \vdash \mathbf{N}(\phi \wedge \psi)$

The second one is a demonstration that agglomeration follows from rules (α) and (β). Since (α) is indisputable, they conclude (β) is invalid.

First, a preliminary definition. The definition I am going to present is provided by Pruss (2013) in order to capture formally the more intuitive notion of "no-choice-about"(see also Huemer 2000, Carlson 2000).

Definition 2.1. $\mathbf{N}\phi$ if and only if $\phi \wedge \sim\exists x \exists \alpha [Can(x, \alpha) \wedge (Does(x, \alpha) \boxdot\!\!\rightarrow \sim\phi)]$

In the definition above, $\boxdot\!\!\rightarrow$ is the subjunctive conditional or counterfactual, x is a variable ranging over humans, α is a variable ranging over all possible action types in the past, present, and future (cf. Pruss (2013: 433)), while $Can(x, \alpha)$ and $Does(x, \alpha)$ are left unanalysed.

The intended interpretation is supposed to capture the idea that $\mathbf{N}\phi$ is incompatible with any humans having free will. Let P be any true proposition about human action. If no one can do anything such that, if one were to do it, $\sim P$ would be the case, then there's no free will.

Though $Can(x, \alpha)$ is left unanalysed, the proofs I will present work given just some fairly plausible assumption about it. With respect to $\Box\!\!\rightarrow$, just to refresh the reader's memory, I shall start by saying something about the meaning of counterfactuals that amounts to the common aspects of Lewis and Stalnaker approaches:

(LS) $\phi \Box\!\!\rightarrow \psi$ is true in a world w if and only if ψ is true in all the worlds in which ϕ is true that are closest to w.

I also adopt the standard terminology in saying that an ϕ-world is just a world in which ϕ is true. In this sense to say that $\phi \Box\!\!\rightarrow \psi$ is true in w is to say that ψ is true in all the ϕ-worlds closest to w.

Now the counterexample. Imagine a situation in which we have a fair coin, and suppose that the coin is not tossed, though surely someone could have tossed it. Let P stand for the proposition expressed by the sentence "the coin does not land heads" and Q for "the coin does not land tails". The counterexample goes as follows. No one has (or ever had) any choice about whether P is true, and the same goes for Q. But someone does have a choice about whether $P \wedge Q$ is true.

According to our interpretation, to say that no one has (or ever had) any choice about whether P is true is to say that P is true and there's no agent x, action-type α such that, if x were to do α, $\sim P$ would be the case. P is true in that scenario because the coin is not tossed, so that the coin does not land heads. (And the same goes for Q). But what about the counterfactual? Is there an action that an agent can perform such that, if she were to perform it, the coin would land heads? We're assuming in the scenario that someone could have tossed the coin. So, the main point is whether at least one of the following counterfactuals is true:

C1: If someone were to toss the coin, it would land heads.

C2: If someone were to toss the coin, it would land tails.

If both C1 and C2 are false, then the premises of agglomeration are true. The conclusion is false because you can toss the coin, for example. And if you were to do it, the coin would land either heads or tails, so that $P \wedge Q$ would be false. Now, are C1 and C2 really false?

Many people think that they are false. To illustrate this, we can have a look at a Lewisian account of counterfactuals, which counts C1 and C2 as false. According to such a view, in order for either counterfactual to be true, either all of the closest worlds to ours where the coin is tossed are heads-worlds or all of them are tails-worlds. For example, if C1 is true, then there must be some factor that influences the similarity relation so that heads-worlds are closer to our world than tails-worlds. But it seems that there is nothing that would grant some special priority to heads-worlds over tails-worlds (or vice versa). Some of such worlds are heads-worlds, while others are tails-worlds...

We are assuming in the scenario that it cannot be the case that the coin lands heads *and* tails. In this sense, we can think of the disjunction "either C1 or C2" as an instance of the conditional excluded middle principle: $(P \mathrel{\Box\!\!\rightarrow} Q) \vee (P \mathrel{\Box\!\!\rightarrow} {\sim}Q)$. As we all know, the principle fails on Lewis' view. The principle fails because Q might be true in some (but not all) of the closest worlds where P is true *and* $\sim Q$ might be true in some (but not all) of the closest possible worlds where P is true. In this case, both $P \mathrel{\Box\!\!\rightarrow} Q$ and $P \mathrel{\Box\!\!\rightarrow} {\sim}Q$ are false.

What's true on Lewis' account is the corresponding "might-counterfactual", that is, "if ϕ were the case, ψ might be the case". Let $\Diamond\!\!\rightarrow$ stand for *if... might....* Now define $\Diamond\!\!\rightarrow$ thus:

(Duality): $\phi \mathrel{\Diamond\!\!\rightarrow} \psi$ if and only if ${\sim}(\phi \mathrel{\Box\!\!\rightarrow} {\sim}\psi)$

What is true on Lewis' view is that, if someone were to toss the coin, it *might* land heads; and if someone were to toss the coin, it *might* land tails. But this

means that, given the definition of the might-counterfactual, C1 and C2 are false. And if they are false, the premises of McKay and Johnson's counterexample are true.

That was the first step of the counterexample. The second step is simpler, and it can be spelled out as follows:

1	$\mathbf{N}P$	premise
2	$\mathbf{N}Q$	premise
3	$\Box(P \supset (Q \supset (P \wedge Q)))$	Logical truth
4	$\mathbf{N}(P \supset (Q \supset (P \wedge Q)))$	α, 3
5	$\mathbf{N}(Q \supset (P \wedge Q))$	β, 1, 4
6	$\mathbf{N}(P \wedge Q)$	β, 2, 5

As we already saw, the premises are true. Rule (α) is valid. But the conclusion is false. Therefore, in the reasoning above, what allowed us getting a false conclusion from true premises was precisely rule (β). Therefore, (β) is invalid.

I think the counterexample is perfectly convincing if we assume a Lewisian theory of counterfactuals. I say "Lewisian" because the first step of the counterexample need not be backed by Lewis' own account of counterfactuals, but merely by the claim that C1 and C2 are both false, and consequently by the denial of the conditional excluded middle principle. If, on the other hand, we were to accept the principle, then the first step would not get off the ground. In order to illustrate this, I will show how things are different if we accept Stalnaker's view, one that accepts the conditional excluded middle principle.

3 Stalnaker's view and the limit assumption

Contrary to Lewis, Stalnaker thinks that there is never more than one closest ϕ-world. This is the limit assumption. Accepting the limit assumption has

important consequences on the exact formulation of truth conditions for counterfactuals as well as on the deductive rules for the logic of counterfactuals. Here I will mention two of them. They will allow us to demonstrate that both agglomeration and (β) hold on Stalnaker's view. Here's the first one (see also Bonevac 2003: 418).

(S) $\sim(\phi \boxrightarrow \psi) \vdash \phi \boxrightarrow \sim\psi$

It's easy to see that (S) holds on Stalnaker's theory. It follows from what has been said about the meaning of counterfactuals (namely, LS) *and* the limit assumption, that there is never more than one closest ϕ-world. (On Lewis' view, on the other hand, (S) doesn't hold. What follows from $\sim(\phi \boxrightarrow \psi)$ is the might-counterfactual $\phi \diamond\!\!\rightarrow \sim\psi$).

(S) goes in only one direction, so that $\sim(\phi \boxrightarrow \psi)$ does not follow from $\phi \boxrightarrow \sim\psi$, since $\phi \boxrightarrow \sim\psi$ and $\phi \boxrightarrow \psi$ hold in case ϕ is impossible. But it does follow if we suppose that $\diamond\phi$.

(S-2) $\diamond\phi, \phi \boxrightarrow \sim\psi \vdash \sim(\phi \boxrightarrow \psi)$

To show that agglomeration holds we also need the assumption that if someone can perform some α, then it is possible she performs α: in other words, I shall assume that $Can(x, \alpha)$ entails $\diamond Does(x, \alpha)$. This assumption isn't wholly unproblematic nowadays because Spencer (2017) has recently provided a battery of cases against this principle. But I don't think that this is really relevant for this discussion. The reason is that what motivates the assumption is rule (α) (which, by the way, is invalid if Spencer is right). However, a crucial aspect of McKay and Johnson's argument is that (α) is valid, as it is explicit in the second step. So if the assumption doesn't work, neither does McKay and Johnson's counterexample to the original rule (β).

Here's the proof of agglomeration on Stalnaker's theory:

1	$\mathbf{N}\phi$	
2	$\mathbf{N}\psi$	
3	$\sim\mathbf{N}(\phi \wedge \psi)$	
4	$\phi \wedge \sim\exists x \exists \alpha[Can(x,\alpha) \wedge (Does(x,\alpha) \boxrightarrow \sim\phi)]$	def. N, 1
5	$\psi \wedge \sim\exists x \exists \alpha[Can(x,\alpha) \wedge (Does(x,\alpha) \boxrightarrow \sim\psi)]$	def. N, 2
6	$\sim[(\phi \wedge \psi) \wedge \sim\exists x \exists \alpha[Can(x,\alpha) \wedge (Does(x,\alpha) \boxrightarrow \sim(\phi \wedge \psi)]$	def. N, 3
7	$\sim(\phi \wedge \psi) \vee \exists x \exists \alpha[Can(x,\alpha) \wedge (Does(x,\alpha) \boxrightarrow \sim(\phi \wedge \psi))]$	Taut con, 6
8	ϕ	$\wedge$E, 4
9	ψ	$\wedge$E, 5
10	$\phi \wedge \psi$	$\wedge$I, 8, 9
11	$\exists x \exists \alpha[Can(x,\alpha) \wedge (Does(x,\alpha) \boxrightarrow \sim(\phi \wedge \psi))]$	Taut con, 7, 10
12	$Can(s,a) \wedge (Does(s,a) \boxrightarrow \sim(\phi \wedge \psi))$	$\exists$E, 11
13	$\sim\exists x \exists \alpha[Can(x,\alpha) \wedge (Does(x,\alpha) \boxrightarrow \sim\phi)]$	$\wedge$E, 4
14	$\sim\exists x \exists \alpha[Can(x,\alpha) \wedge (Does(x,\alpha) \boxrightarrow \sim\psi)]$	$\wedge$E, 5
15	$\sim Can(s,a) \vee \sim(Does(s,a) \boxrightarrow \sim\phi)$	Taut con, 13
16	$\sim Can(s,a) \vee \sim(Does(s,a) \boxrightarrow \sim\psi)$	Taut con, 14
17	$Can(s,a)$	$\wedge$E, 12
18	$\sim(Does(s,a) \boxrightarrow \sim\phi)$	DS, 15, 17
19	$\sim(Does(s,a) \boxrightarrow \sim\psi)$	DS, 16, 17
20	$Does(s,a) \boxrightarrow \phi$	S, 18
21	$Does(s,a) \boxrightarrow \psi$	S, 19
22	$Does(s,a) \boxrightarrow \sim(\phi \wedge \psi)$	$\wedge$E, 12
23	$Does(s,a) \boxrightarrow (\phi \wedge \psi)$	Taut con, 20, 21
24	$\Diamond Does(s,a)$	Can, 17
25	$\sim(Does(s,a) \boxrightarrow (\phi \wedge \psi))$	S-2, 22, 24
26	$\bot$	
27	$\mathbf{N}(\phi \wedge \psi)$	$\neg$I, 3–26

The upshot is that agglomeration holds on Stalnaker's theory given a fairly plausible assumption about "can". This shouldn't be too surprising because the conditional excluded middle holds on Stalnaker's view, so that C1 and C2 aren't both false.

"So much worse for Stalnaker's theory!", one might reply. Well, indeed, if Stalnaker's view counts C1 and C2 as true, then we might have a problem. But Stalnaker's theory doesn't count C1 and C2 as true. He counts them as neither true nor false.

> This time someone ran off with the coin before it was tossed. Having no other coin, Tweedledee and Tweedledum argue about how it would have landed if it had been flipped. Tweedledee is convinced that it would have landed heads, Tweedledum that it would have landed tails. Again, neither has a reason – they agree that the coin was a normal one and that the toss would have been fair. This time, there is little inclination to say that one of them must be right. Unless there is a story to be told about a fact that renders one or the other of the counterfactuals true, we will say that neither is. (Stalnaker 1984: 165)

What Stalnaker does is to combine his account with the theory of *supervaluations* (Stalnaker 1980: 90). This is why he takes the truth-values of these counterfactuals to be indeterminate, that is, neither true nor false. In other words, Stalnaker's theory also produces the desired result that C1 and C2 are not true. But rather than saying they are false, he takes their truth-value as indeterminate.

On the standard account of supervaluationism, a sentence is true if it is true on all precisifications, false if it is false on all precisifications, and neither true nor false otherwise. C1 and C2 are neither true nor false on all precisifications.

Thus they are neither true nor false. If truth is truth on all precisifications then supervaluationists account for validity in the following way: an argument is globally valid if and only if if the premises are true on all precisifications the conclusion is true on all precisifications. Since on Stalnaker's theory the premises of agglomeration are not true on all precisifications, we cannot say agglomeration is invalid. So, if we are sympathetic to Stalnaker's theory, we cannot assume that the counterexample is successful. The counterexample does not show a situation in which the premises are true and the conclusion false. It shows instead a situation in which the premises are indeterminate and the conclusion is false.

This result is totally in line with the premises of McKay and Johnson's counterexample being indeterminate rather than true, so that there is no situation in which the premises of agglomeration are true and the conclusion is false. But if agglomeration holds, then one need not be worried about McKay and Johnson's argument after all. In fact, the second interesting result is that the original rule (β) holds on Stalnaker's theory.

1	$\mathbf{N}\phi$	
2	$\mathbf{N}(\phi \supset \psi)$	
3	$\phi \wedge {\sim}\exists x \exists \alpha[Can(x,\alpha) \wedge (Does(x,\alpha) \boxrightarrow {\sim}\phi)]$	def. N, 1
4	$(\phi \supset \psi) \wedge {\sim}\exists x \exists \alpha[Can(x,\alpha) \wedge (Does(x,\alpha) \boxrightarrow {\sim}(\phi \supset \psi))]$	def. N, 2
5	ϕ	$\wedge$E, 3
6	$\phi \supset \psi$	$\wedge$E, 4
7	ψ	$\Rightarrow$E, 5, 6
8	$Can(s,a) \wedge (Does(s,a) \boxrightarrow {\sim}\psi)$	
9	$Can(s,a)$	$\wedge$E, 8
10	$Does(s,a) \boxrightarrow {\sim}\psi$	$\wedge$E, 8
11	$\Diamond Does(s,a)$	Can, 9
12	${\sim}\exists x \exists \alpha[Can(x,\alpha) \wedge (Does(x,\alpha) \boxrightarrow {\sim}\phi)]$	$\wedge$E, 3
13	${\sim}\exists x \exists \alpha[Can(x,\alpha) \wedge (Does(x,\alpha) \boxrightarrow {\sim}(\phi \supset \psi))]$	$\wedge$E, 4
14	$\forall x \forall \alpha {\sim}[Can(x,\alpha) \wedge (Does(x,\alpha) \boxrightarrow {\sim}\phi)]$	Taut con, 12
15	${\sim}Can(s,a) \vee {\sim}(Does(s,a) \boxrightarrow {\sim}\phi)$	$\forall$E, 14
16	${\sim}Can(s,a) \vee {\sim}(Does(s,a) \boxrightarrow {\sim}(\phi \supset \psi)$	$\forall$E, 13
17	${\sim}(Does(s,a) \boxrightarrow {\sim}\phi)$	DS, 9, 15
18	${\sim}(Does(s,a) \boxrightarrow {\sim}(\phi \supset \psi)$	DS, 9, 16
19	$Does(s,a) \boxrightarrow \phi$	S, 17
20	$Does(s,a) \boxrightarrow (\phi \supset \psi)$	S, 18
21	$Does(s,a) \boxrightarrow \psi$	Taut con, 19, 20
22	${\sim}(Does(s,a) \boxrightarrow \psi)$	S-2, 10, 11
23	$\bot$	
24	${\sim}[Can(s,a) \wedge (Does(s,a) \boxrightarrow {\sim}\psi)]$	$\neg$I, 8–23
25	$\forall x \forall \alpha {\sim}[Can(s,a) \wedge (Does(s,a) \boxrightarrow {\sim}\psi)]$	$\forall$I, 24
26	$\psi \wedge \forall x \forall \alpha {\sim}[Can(s,a) \wedge (Does(s,a) \boxrightarrow {\sim}\psi)]$	$\wedge$I, 7, 26
27	$\mathbf{N}\psi$	def. N, 26

Stalnaker gives a theory of counterfactuals that differs over details. But the details matter because they are about how we should express the negation of a counterfactual conditional. If the negation of a counterfactual $\phi \mathrel{\Box\!\!\rightarrow} \psi$ is $\phi \mathrel{\Box\!\!\rightarrow} {\sim}\psi$, and ${\sim}(\phi \mathrel{\Box\!\!\rightarrow} \psi)$ is also a negation of the former, then $\phi \mathrel{\Box\!\!\rightarrow} {\sim}\psi$ is equivalent to ${\sim}(\phi \mathrel{\Box\!\!\rightarrow} \psi)$. For conditionals with possibly true antecedents, this is equivalent to the conditional excluded middle: $(\phi \mathrel{\Box\!\!\rightarrow} \psi) \vee (\phi \mathrel{\Box\!\!\rightarrow} {\sim}\psi)$. So the proof really depends only on the conditional excluded middle principle, as it relies on the deductive rules (S) and (S-2). It is an important detail after all, since it affects the validity of the main argument for incompatibilism.

So far, so good. But now one may fairly ask: how relevant is this result in the context of the current debate concerning the consequence argument? The simplest answer would be to accept Definition 2.1 and the limit assumption and run the consequence argument with the **N** operator. Yet this might not be something that the incompatibilist advocates of the argument would be happy to accept, since they would rather not take a stand on counterfactuals. Instead, they would appeal to a revision of the argument in order to put forward a version of the consequence argument which is valid regardless of controversial assumptions about counterfactuals, such as the conditional excluded middle. However, I shall argue that this move has a cost. The cost is that there is some particular line of argument against the premises of the revised versions of the consequence argument that does not succeed against the original version.

4 Revising the consequence argument

One important incompatibilist response to McKay and Johnson's counterexample to (β) is to accept the counterexample and adopt a different meaning of "no-choice about" in terms of the might-counterfactual (O'Connor 2000; Carlson 2000; Huemer 2000; Pettit 2002; Turner 2009; and Steward 2012). That

is, even though there is nothing I can do such that, if I were to do it, the coin would land heads (tails), there is something I can do such that, if I were to do it, the coin *might* land heads (tails).

We can thus introduce the following operator:

Definition 4.1. $\mathbf{M}\phi$ if and only if $\phi \wedge \sim\exists x \exists \alpha [Can(x, \alpha) \wedge (Does(x, \alpha) \diamond\!\!\rightarrow \sim\phi)]$

The solution then is to run the consequence argument using the **M** operator, which is linked to abilities and might-counterfactuals in the way that the **N** operator is linked to abilities and counterfactuals. Accordingly, the corresponding (β)-rule will not be open to McKay & Johnson's counterexample (Carlson 2000: 286–87):

(β-M): $\mathbf{M}\phi$, $\mathbf{M}(\phi \supset \psi) \vdash \mathbf{M}\psi$

This yields, to be sure, a valid formulation of the consequence argument, regardless of whether or not we accept the limit assumption.

Notice that the premises will now have to be formulated with **M**. $\mathbf{M}L$, for example, says that no one can do anything such that L *might* be false, which is different from saying that L *would* be false. How different? That depends on how we understand the might-counterfactual. While we can all agree that "If someone tossed the coin, it might have come down heads" and "If someone tossed the coin, it might have come down tails" are both true, there is an important dispute among counterfactual theorists over *why* the might-counterfactual is true. Why is it true that "If someone tossed the coin, it might have come down heads (tails)"?

At a first glance, one could say, following Lewis, that $P \diamond\!\!\rightarrow Q$ is true just because $\sim(P \mathrel{\Box\!\!\rightarrow} \sim Q)$ is.[3] But many counterfactual theorists find this troublesome because it treats "if... might" as if it were an idiom, as Bennett says,

[3]Of course, Stalnaker (1981) would demur. On his approach might-counterfactuals are

"something to be understood as a single linguistic lump, like 'under way': you wouldn't try to explain 'At 5 p.m. the ship got under way' by explaining 'under' and explaining 'way"' (2003: 189). The alternative and shared view is that the might-counterfactual could be analysed in terms of the separate meanings of "if" and "might" (Stalnaker 1981; Bennett 2003; Gillies 2010; Williams 2010). Moreover, the **M**-version of the consequence argument should go through for any plausible account of might-counterfactuals, or at least for any account which accommodates McKay & Johnson's counterexample, and should not take a stand on a specific view such as Lewis' and his idiomatic treatment of "if... might...". That again would force the incompatibilist to take a stand on counterfactuals.

But now consider the view that a might-counterfactual is a counterfactual with a modal operator expressed by "might" as $\blacklozenge$ embedded in its consequent, so that the logical form of the might-counterfactual is $P \mathrel{\Box\!\!\rightarrow} \blacklozenge Q$ (Bennett 2003; Gillies 2010; Goldstein 2020).[4] How $\blacklozenge$ is understood is a matter of controversy[5], but Bennett's view – which is line with McKay & Johnson's counterexample – is that $\blacklozenge P$ means that P is compatible with knowledge of all the present facts. To get a gist of it let us have a look at our paradigm case. Let T be "the coin is tossed" and H "the coin lands heads". The sense in which $T \mathrel{\Box\!\!\rightarrow} \blacklozenge H$ is true is the sense in which even if we had complete knowledge of everything

analysed as an epistemic possibility claim at wide scope to a subjunctive conditional, such as $\Diamond_e(P \mathrel{\Box\!\!\rightarrow} Q)$, where $\Diamond_e$ stands for epistemic possibility. Lewis objects to this view (Lewis 1973: 80), but see De Rose (1994) for an intricate response.

[4]Another proposal is that $\blacklozenge$ has wide scope, so that the logical form of the might-counterfactual is $\blacklozenge(P \mathrel{\Box\!\!\rightarrow} Q)$. This is the view of Stalnaker (1981) and De Rose (1991), but since they both accept the conditional excluded middle, the consequence argument could be ran with **N** on their approach.

[5]The dominant view as to "might" is that it expresses some sort of epistemic possibility, something (not quite but) along these lines: "might P" is true (as uttered by x) iff P is compatible with what x knows (Williams 2010: 655); De Rose's view (1991) is that "might P" is only true when we do not know nor could in relevant ways find out anything incompatible with P. Interestingly, if anything along these lines is correct, it does not matter what the scope of $\blacklozenge$ is, for $\mathbf{M}L$ and $\mathbf{M}P_0$ will be too difficult for the incompatibilist to defend. For instance, $\mathbf{M}L$ could be objected by simply pointing out that $\sim L$ is consistent with everything we know or that the conditional "If I were to do otherwise, $\sim L$" is consistent with everything we know. This seems quite plausible because surely the incompatibilist will not want to be committed to the claim that "might $\sim L$" is false according to the epistemic "might".

that is true by the time T is true, that would not be sufficient for determining whether H is true (because the process is, say, indeterministic). In other words, H is compatible with what an ideal agent knows by the time the coin is tossed, which includes a complete list of everything that is true just before H is true. Or: Nothing is the case that rules H out (2003: 190).

Of course, given this account, anyone attracted to a Humean view of the laws will find little reason to accept $\mathbf{M}L$. This sort of objection is well known in the literature (Beebee 2000, 2003; Gustaffson 2017), but there are some novel points worth making. Let us first draw a picture of the objection in order to facilitate discussion. Suppose we have a list of everything that is true in our universe up to now. According to Humeans, such a list would not display the laws of nature because the laws are not present in the current state of the universe. So not even an ideal knower could know whether L is true, even if she knew everything that is true about the present. The ideal knower would know, to be sure, everything about the current state of our universe, including true generalisations about what has happened; but the current state of the universe *per se* would not, as it were, contain the laws of nature. As a result, it is trivial that any action one is able to perform is such that L might not be the case, where "might" stands for Bennett's understanding of the diamond.

Does the trivial objection hinge on Bennett's view? No. Gustaffson puts forward the Humean objection to $\mathbf{M}L$ based on the intuitiveness of the following principle: "if it's still contingent shortly after P is made true whether $\sim Q$ will be made true, then the 'might' counterfactual 'If it were the case that P, it might be the case that $\sim Q$' is true, even if P and Q are true (2017: 710). Finch & Warfield (1998) also make use of a similar intuition by arguing that, in cases of indeterministic causation where the fact that DB indeterministically causes the fact that R, "any action (including inaction) at all that one performed is

such that it might have resulted in R's not following DB" (1998: 526). But Bennett's view seems to nicely accommodate these intuitions.

It could be objected that while the Humean view of the laws gives reason for denying $\mathbf{M}L$, it also gives reason for denying $\mathbf{N}L$. Influential compatibilists such as Beebee, for instance, deny both $\mathbf{M}L$ and $\mathbf{N}L$.[6]

Yet it is important to point out that Beebee does not deny $\mathbf{N}L$ based on the trivial objection as above. Beebee argues against $\mathbf{N}L$ based on the *standard* best system account of the laws (BSA) – according to which the laws of nature are those generalisations that best combine simplicity and strength – *and* a Humean ontology. Beebee is correct, I think, in saying that *if* we accept the standard BSA and the Humean picture when it comes to fundamental ontology, then there seems to be little reason to accept $\mathbf{N}L$. But this is mainly because, back in the day, the standard BSA, in and of itself, did not have the resources to account for L's counterfactual resilience. If the laws are just efficient summaries of particular matters of fact, and the facts are void of modality, it seems obvious that if the facts were different, so would be the laws. This makes room for free will, but at the cost of giving up certain intuitions concerning the laws of nature. In particular, the intuition that L *would* still be true if the particular matters of fact were different. Consider what Demarest (2017) calls the "impoverished worlds objection".

Let L stand for "all massive particles attract each other". Since L is a law of nature, as philosophers, we might well wonder whether L *would* still be a

[6] Vihvelin (2013) is another influential compatibilist who denies $\mathbf{N}L$. But she denies $\mathbf{N}L$ because she thinks – incorrectly in my view – that Lewis' theory of counterfactuals is enough to deny $\mathbf{N}L$, regardless of whether or not we accept a Humean view of the laws (2013: 163). I agree that if Lewis' theory of counterfactuals is correct, then $\mathbf{N}L$ is false. But his theory clearly hinges on a Humean view of the laws. One distinctive feature of the *governing* theories is that the laws are counterfactually resilient, that is, the laws of nature would still be true in counterfactual suppositions logically consistent with them. This "Nomic Preservation" principle or something like it is accepted by mainstream governing views of the laws and is inconsistent with the denial of $\mathbf{N}L$. See, for instance, Lange (2000, 2009), Maudlin (2007), Goodman (1983), Carroll (1994), and Roberts (2008).

law if there was a world with only one massive particle travelling inertially in perpetuity. Intuitively, or so the objector says, L would still be a law in that world; because it seems that if there were another particle there, they would attract each other. However, according to the standard BSA, L would not be a law. If there is a possible world with just one massive particle travelling inertially, then it will be a law that "all massive particles travel inertially". It intuitively delivers the wrong result.

Humeans have replied to the objection in a number of ways[7], but it seems as though the Humean view of the laws is at odds with our pre-theoretical intuitions concerning the laws and counterfactuals. So it does not come as a surprise that $\mathbf{N}L$ can also be denied if such a view is assumed.

Nevertheless, thanks to some recent developments of the best system accounts, we need neither accept a Humean ontology nor the standard BSA to run the trivial objection to $\mathbf{M}L$. I have in mind some contemporary developments in the laws of nature literature, such as the potency-BSA (Demarest 2017) and the best predictive system account of the laws (Dorst 2018). Both views can account for L's counterfactual resilience, namely, that L would still be true in counterfactual assumptions consistent with it.[8] On these views, the trivial objection to $\mathbf{M}L$ will not affect $\mathbf{N}L$, for even though L may be false relative to idealised epistemic possibility (because knowledge of all the present facts may not be enough to know what the laws are), the laws will still support their counterfactuals, so that L will not be false in the closest worlds to ours.

Demarest promotes a scientific package that is Humean when it comes to the laws of nature but anti-Humean concerning the fundamental properties; namely, dispositional essentialism, the view that at least some properties at the

[7] Carroll (1994) offers another version of that argument, which is interestingly responded by Beebee (2000).

[8] See, for instance, Dorst (2020) with respect to the predictive BSA and Demarest (2017) with respect to the potency BSA.

fundamental level are potencies, that is, properties with dispositional essences (Ellis 2001; Bird 2007; Jacobs 2010; Tugby 2013). The dispositionalist will tell us, say, that anything with the dispositional property of positive charge is essentially disposed (or has the power) to attract negative charges and repel positive charges. On Demarest's view, the laws are those axioms that best systematise all of the possible distributions of the fundamental dispositional properties. Here is how she deals with the impoverished worlds objection:

> Consider, again, a world with a single massive particle, traveling inertially for all time. The laws of this world will systematize not just this world, but all worlds that contain mass. Therefore, it will be a law that all massive particles attract each other, and NOT that they always travel inertially. (2017: 51).

The view then allegedly accommodates L's counterfactual resilience. Still, there is no guarantee – nor should there be – that $\sim\blacklozenge\sim L$. Even assuming we have complete knowledge of the present state of the universe, that may not be enough to identify every possible distribution of the potencies. The core doctrine of dispositional essentialism is that the powers of a property are essential to it; it does not claim anything about *when* all the properties can come to be distributed and instantiated in our world.[9]

All of this was to point out that compatibilists can be more concessive than Beebee and agree with incompatibilists that the laws of nature would not have been false had someone acted otherwise. Perhaps most incompatibilist were not that worried about the standard Humean objection to $\mathbf{N}L$ because it presupposed some sort of revisionist view of the laws. After all, on Beebee's view the laws of nature are *not* counterfactually resilient with respect to human actions.

[9] What is more, Tugby (2016) interestingly argues that it is far from clear how dispositional essentialists can explain that our world is regular rather than chaotic from moment to moment. An ideal knower could know a great deal of regularities up to now, but then if our world starts to be chaotic, the axioms of our best system will be altogether different from those regularities.

But as far as I can see, compatibilists may concede that $\mathbf{N}L$ is true without conceding that $\mathbf{M}L$ is true too. The upshot is that we have a particular line of argument against the **M**-version of the consequence argument that does not succeed against the **N**-version.

Interestingly, the very same line of objection will also affect another influential incompatibilist line of response to McKay & Johnson. According to this sort of response, the incompatibilist need not worry about defending $\mathbf{M}L$ because she can put forward the consequence argument with a different rule, such as

($\beta - 2$): $\mathbf{N}\phi, \Box(\phi \supset \psi) \vdash \mathbf{N}\psi$.

($\beta - 2$) was originally proposed by David Widerker (1987) and then defended by Alicia Finch & Ted Warfield (1998). More recently, Alexander Pruss (2013) has proved that ($\beta - 2$) holds given the weakening principle about counterfactuals, which holds both on Lewis' and Stalnaker's theories. With ($\beta - 2$), the incompatibilist may put forward a new version of the consequence argument:

1	$\Box((L \wedge P_0) \supset P)$	determinism
2	$\mathbf{N}(L \wedge P_0)$	premise
3	$\mathbf{N}P$	$\beta - 2$, 1, 2

Even so, the ($\beta - 2$) version of the consequence argument is not without problems. As Warfield & Finch point out (1998: 523), $\mathbf{N}(L \wedge P_0)$ is formally stronger than the premises $\mathbf{N}L$ and $\mathbf{N}P_0$. Even if the original premises were true, that alone would not be enough to conclude that $\mathbf{N}(P_0 \& L)$ is true too, since the lesson of McKay & Johnson's argument – if it has any bite – is that **N** is *not* agglomerative. So the opponent of the consequence argument may now fairly ask why she should accept a premise that is formally stronger than those in the original version.

In light with this conundrum, Warfield & Finch offer an original defence of

$\mathbf{N}(L \wedge P_0)$. As they put it, $L \wedge P_0$ "offers a description of what might be called the "broad past" – the complete state of the world at a time in the distant past including the laws of nature" (1998: 523). If the broad past is fixed in the way that the remote past is fixed, then it seems premise 2 is true. But why should we think that the broad past is fixed? According to the authors:

> We think that the claims that the laws are inalterable and do not change combined with the point that this implies that they are, in a sense, a part of the past, is a plausible explanation of this intuition about their fixity (1998: 523, footnote 15).

What is there to say in response? Finch & Warfield talks as though the laws of nature were necessitation relations among universals (Armstrong 1983; Dretske 1977; Tooley 1987), so that the laws would already be present in the distant past. Clearly, however, such a defence of premise 2 does not hold water according to systematising accounts of the laws, such as those discussed before. This is because the laws will not be part of the past, and on these views there is no broad past to get the incompatibilist argument off the ground. Given this much, the only available defence of premise 2 is quite unpersuasive for a wide range of contemporary systematising theories of the laws.[10] $\mathbf{N}(L \wedge P_0)$ is left undefended.

[10] I regret that I do not know whether systematising views can or should account for the counterfactual resilience of the *broad* past; plausibly there is no broad past on systematising views. Should the broad past be fixed in the impoverished worlds scenario? If it is a law that all massive particles attract each other in that world, and the broad past is fixed, then the counterfactual "if there were two particles, they would attract each other" is a counterpossible; but then it would also be true that "if there were two particles, they would not attract each other". But it does not seem Demarest would count them as counterpossibles. Perhaps what is really a threat to free will on a view such as Demarest is not the deterministic laws, but the causal powers of the fundamental properties.

5 Final remarks

All in all, the revised formulations of the consequence argument are not without problems. This is not to say that they are not important, for they do the nice job of avoiding McKay & Johnson's objection to the consequence argument. In this sense, they are just fine, because they do indeed show that there are seemingly valid formulations of the consequence argument.[11] But if the incompatibilist wants to get rid of Humean criticisms, adopting the revised versions is not the way to go. This is why the approach proposed in this paper is important, for we can maintain the original deduction rules and the original premises of the consequence argument while avoiding the Humean criticism.[12] Of course, this move has significant costs as well (e.g. Stalnaker's view of counterfactuals, supervaluationism, the assumption about "can", etc.). But as a reviewer pointed out, perhaps we need not be forced to choose between the standard revisions of the consequence argument and the approach that makes use of a particular theory of counterfactuals. It might just be that different approaches are more successful in some aspects but less successful in other aspects.

To sum up: I have argued that there are two important lessons of McKay & Johnson's objection that have been overlooked in the literature. The first is that the objection presupposes the falsity of the conditional excluded middle, so that

[11] That is, provided we assume the counterfactual sufficiency interpretation of the **N** operator, which may be properly challenged. See, for instance, Lampert and Merlussi (2021, forthcoming).

[12] Moreover, as an anonymous reviewer pointed out, we are also able to preserve the similarities between the *Mind* argument and the consequence argument. For example, if we interpret the relevant counterfactuals in terms of Stalnaker's theory, we could run the *Mind* argument with the **N** operator. This might be useful if someone wants to defend the view that free will is metaphysically impossible (because $\mathbf{N}p$ would be incompatible with determinism and indeterminism) or perhaps van Inwagen's mysterianism. My worry, however, is that the counterfactual interpretation of **N** may fail to properly capture the control condition related to free will. See, for instance, Lampert and Merlussi (2021, forthcoming) for arguments against the counterfactual sufficiency interpretation of **N**. In fact, the arguments from Lampert and Merlussi, if successful, would affect any common formulation of the consequence argument using either of (α), (β), or $(\beta - 2)$, regardless of whether counterfactuals are interpreted with a Lewisian or an alternative, Stalnakerian semantics.

the counterexample to rule (β) is not as robust as we used to think. In order to make this point I demonstrated that the original rule is valid on a different, well motivated theory of counterfactuals; Stalnaker's theory. Second, the revised versions of the consequence argument are threatened by recent systematising views of the laws, but these views do not threaten the approach proposed in this paper.

References

Armstrong, D., 1983. *What is a Law of Nature?*, Cambridge: Cambridge University Press.

Beebee, H., 2000. "The Nongoverning Conception of Laws of Nature", *Philosophy and Phenomenological Research* 61: 571–594.

Beebee, H., 2002. "Reply to Huemer on the Consequence Argument", *The Philosophical Review* 111 (2): 235–241.

Beebee, H., 2003. "Local Miracle Compatibilism", *Noûs* 37 (2): 258–277.

Bennett, J., 2003. *A philosophical guide to conditionals*, Oxford: Oxford University Press.

Bird, A., 2007. *Nature's Metaphysics: Laws and Properties*, Oxford: Oxford University Press.

Blum, A., 2000. "N", *Analysis* 60 (3): 284-286.

Bonevac, D., 2003. *Deduction: Introductory Symbolic Logic*, 2nd edition. Oxford: Blackwell Publishing.

Campbell, J. K., 2007. "Free Will and the Necessity of the Past", *Analysis* 67 (2): 105–111.

Carlson, E., 2000. "Incompatibilism and the transfer of power necessity", *Noûs* 34: 277–290.

Carroll, J., 1994. *Laws of Nature*, Cambridge: Cambridge University Press.

Demarest, H., 2017. "Powerful Properties, Powerless Laws". In J. Jacobs (ed.), *Putting Powers To Work: Causal Powers In Contemporary Metaphysics*, Oxford: Oxford University Press: 39-54.

DeRose, K., 1991. "Epistemic possibilities", *The philosophical review* 100: 581–605.

DeRose, K., 1994. "Lewis on 'might' and 'would' counterfactual conditionals". *Canadian Journal of Philosophy* 24(3): 413–418.

Dorst, C., 2018. "Towards a best predictive system account of laws of nature", *British Journal for the Philosophy of Science* 70: 877–900.

Dorst, C., 2020. "Why do the Laws Support Counterfactuals?", *Erkenntnis*.

Dretske, F. I., 1977. "Laws of Nature", *Philosophy of Science* 44 (2): 248-268.

Finch, A. and Warfield, T., 1998. "The *Mind* argument and libertarianism", *Mind* 107: 515-528.

Gillies, A. S., 2010. "Iffiness", *Semantics and Pragmatics*, 3(4), 1–42.

Goldstein, S., 2020. "The counterfactual direct argument". *Linguist and Philosophy* 43: 193–232

Goodman, N., 1983. *Fact, fiction, and forecast*, Cambridge, MA: Harvard University Press.

Gustafsson, J., 2017. "A strengthening of the consequence argument for incompatibilism", *Analysis* 77: 705-715.

Huemer, M., 2000. "Van Inwagen's Consequence Argument", *Philosophical Review* 109: 525-544.

Jacobs, J., 2010. "A powers theory of modality: Or, how I learned to stop worrying and reject possible worlds", *Philosophical Studies* 151: 227–248.

Lampert, F., and Merlussi, P. (2021). "Counterfactuals, counteractuals, and free choice. *Philosophical Studies*", 178, 445–469.

Lampert, F., and Merlussi, P. (forthcoming). "How (not) to construct worlds with responsibility", *Synthese*.

Lange, M., 2000. *Natural laws in scientific practice*, Oxford: Oxford University Press.

Lange, M., 2009. *Laws and lawmakers*, New York: Oxford University Press.

Lewis, D., 1973. *Counterfactuals*, Cambridge: Harvard University Press.

Lewis, D., 1979. "Counterfactual Dependence and Time's Arrow", *Noûs* 13: 455–76

Lewis, D., 1981. "Are We Free to Break the Laws?", *Theoria* 47: 113 – 21

Lewis, D. 1986. Postscripts to "Counterfactual dependence and time's arrow". In *Philosophical Papers*: Volume II, 52–66. Oxford: Oxford University Press.

Maudlin, T., 2007. *The metaphysics within physics*. Oxford: Oxford University Press.

McKay, T. J. and Johnson, D. 1996. "A Reconsideration of an Argument against Compatibilism", *Philosophical Topics* 24 (2): 113–122.

Mele, A. R., 2006. *Free will and luck*, New York: Oxford University Press.

O'Connor, T., 1993a. "On the transfer of necessity", *Nous* 27: 204-18.

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

Pettit, G. 2002. "Are we rarely free? A response to restrictivism", *Philosophical Studies* 107: 219–37.

Pruss, A. R. 2013. "Incompatibilism Proved", *Canadian Journal of Philosophy* 43: 430–437.

Roberts, J., 2008. T*he law-governed universe*, New York: Oxford University Press.

Spencer, R., 2017. "Able to do the impossible", *Mind* 126: 465-497.

Stalnaker, R., 1968. "A Theory of Conditionals", in *Studies in Logical Theory*, ed. Rescher, N., Oxford: Blackwell.

Stalnaker, R., 1981. "A Defense of Conditional Excluded Middle ", in *Ifs*, Harper W., Stalnaker, R. and Pearce, G., Dordrecht: Reidel.

Stalnaker, R., 1984. *Inquiry*, Cambridge, MA: Bradford Books.

Steward, H., 2012.*A Metaphysics of Freedom*, Oxford: Oxford University Press.

Tooley, M., 1987. *Causation: A Realist Approach*, Oxford: Clarendon Press.

Tugby, M., 2013. "Platonic dispositionalism", *Mind* 122: 451–480.

Tugby, M., 2016. "The problem of retention", *Synthese* 194: 2053–2075.

Turner, J. 2009. "The incompatibility of free will and naturalism", *Australian Journal of Philosophy* 87: 565–87.

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

Van Inwagen, P., 2008a. "The consequence argument". In P. van Inwagen and D. Zimmerman (Eds.), *Metaphysics. The big questions* (2nd ed., pp. 450–456). Oxford: Blackwell.

Vihvelin, K., 2013. *Laws, causes and free will: why determinism doesn't matter*, Oxford: Oxford University Press.

Vihvelin, K., 2017. "Arguments for incompatibilism", Retrieved from: https://plato.stanford.edu/entries/incompatibilism-arguments/.

Widerker, D., 1987. "On an Argument for Incompatibilism" *Analysis* 47 (1): 37–41.

Williams, J. R. G., 2010. "Defending conditional excluded middle", *Noûs* 44(4): 650–668.