A Problem for Counterfactual Sufficiency

John William Waldrop

September 2022

1 Introduction

The consequence argument purports to show that determinism is true only if no one has free will.[1] Judgments about whether the argument is sound depend on how one understands locutions of the form 'p and no one can render p false'.[2] The main interpretation on offer appeals to *counterfactual sufficiency*: s can render p false just in case there is something s can do such that, were s to do it, p would be false; otherwise, s cannot render p false. Here I show that, in the context of the consequence argument, this interpretation conflicts with widely endorsed principles governing the logic of counterfactuals.

2 The Consequence Argument

For the version of the consequence argument we are interested in,[3] we let Φ characterize the intrinsic state of the world at some time in the remote past conjoined with a statement of the laws of nature. On determinism, the laws and past states of the world jointly determine the history of the world at any time; so, if determinism is true then Φ entails every truth. We also let Np abbreviate 'p and no one can render p false.' (As a stylistic variant: 'p and no one has a choice about p'.) Now assume that if a truth p is something no one can render false, and p entails some further truth q, then no one can render q false:

$$\textsc{Transfer}\text{: } Np, \Box(p \to q) \vdash Nq$$

This much facilitates a version of the consequence argument.

[1] The classic statement of the consequence argument owes to Peter van Inwagen in his 1983 monograph *An Essay on Free Will*. It bears mentioning that the consequence argument is not an argument for *incompatibilism*—the thesis that determinism incompatible with free will—but rather for the weaker, non-modal thesis that *in fact* either determinism is false or no one has free will. See Campbell (2007) and Cutter (2017) for discussion.

[2] Following others in the literature, in what follows we will be somewhat sloppy in using expressions like 'p', 'q', etc., as well as more complex expressions, in both object position and sentence position. If needs be, one may read objectual occurrences of 'p' as abbreviating singular terms for propositions, like 'the proposition that p'. We will also proceed with relative carelessness when it comes to marking distinctions between use and mention; this is harmless.

[3] For similar arguments, see Widerker (1987), Finch and Warfield (1998), and Pruss (2013).

Let p be any true sentence about human action—say, 'Clarence raises his hand'. On determinism, Φ entails p. It is plausible that neither Clarence nor anyone else has a choice about the laws or about the past, so that no one can render Φ false. We then get that determinism is true only if nobody—including Clarence—has a choice about whether Clarence raises his hand:

(1)	$N\Phi$	Premise
(2)	$\Box(\Phi \rightarrow p)$	Premise from determinism
(3)	Np	From 1 and 2 by TRANFER

There is nothing peculiar about p—the same sort of argument applies to any truth whatsoever. So, if determinism is true no one can render false any truth about human actions, no one can do otherwise than they in fact do, and therefore no one has free will.

3 Counterfactual Sufficiency

Setting aside any concerns about the merits of the premises, the argument is sound only if the rule TRANSFER is itself valid. Is it valid? Granting that we have some grip on the notion of being able to render something false—though the latter notion is admittedly somewhat technical and rarefied—we can investigate this question by supplying a plausible interpretation of the N operator.

The most prominent interpretation on offer is what Michael Huemer (2000) calls *the counterfactual sufficiency interpretation*, according to which being able to render p false amounts to being able to do something such that, were one to do it, p would be false. Accordingly, Np is true if and only if (a) p is true and (b) no one can do anything such that, were they do to do it, p would be false.[4]

This interpretation enjoys considerable support in the literature. It corresponds to Lewis's (1981) characterization of being able to render a proposition false *in the weak sense*. Finch and Warfield (1998: 525) associate the counterfactual sufficiency interpretation with van Inwagen's original interpretation of the N operator; Nelkin (2001) and Baker (2008) appear to follow Finch and Warfield here. Graham also (2010) endorses this interpretation and, like Finch and Warfield, recommends it as a faithful explication of van Inwagen's original interpretation. Still others variously endorse the counterfactual sufficiency interpretation or, without endorsing it, treat it as a viable interpretation of van Inwagen's N operator.[5]

Moreover, as one might have hoped, the counterfactual sufficiency interpretation interfaces fruitfully with rigorous inquiry into whether or not TRANSFER is valid. In a 2013 paper, Pruss shows that in standard logics for counterfactuals

[4]The counterfactual sufficiency interpretation is to be contrasted with its chief rival, according to which Np is true if and only if (a) p is true and (b) no one can do anything such that, were they to do it, p *might* be false. See Huemer (2000), van Inwagen (2000), van Inwagen (2015).

[5]See Carlson (2000), Pruss (2013), and Gustafsson (2017). Additionally, Finch (2012) glosses the related locution 'it is not up to S at t whether p' along similar lines.

one can prove TRANSFER relative to the counterfactual sufficiency interpretation, duly regimented.[6]

So, the counterfactual sufficiency interpretation not only enjoys considerable support, but it is also serviceable for underwriting TRANSFER for the purposes of lodging a valid consequence argument. What is so far lacking is a rigorous defense of the counterfactual sufficiency interpretation itself. We may well ask: Is the counterfactual sufficiency interpretation a plausible or otherwise viable interpretation of the notion of being able to render something false? Given widely endorsed principles about the logic of counterfactuals and standard verdicts about cases in the literature, it appears that the counterfactual sufficiency interpretation is *not* a viable interpretation. This much I shall defend in what follows.

Before that, though, we make the following clarification. In this context, we assume that the counterfactual sufficiency interpretation *does not* simply amount to a stipulation as to the meaning of 'can render false'—introduced as a technical term or as a term of art.[7] Instead, as we have said, we assume that we have some antecedent grip the notion of one's being able to render something false. Otherwise, there would be little point in deriving consequences from candidate interpretations and comparing them with common verdicts about cases, as we will do below. In what follows, then, the counterfactual sufficiency interpretation shall be understood as a substantive, rather than a merely stipulative, claim about the meaning or intension of 'can render false'.[8]

4 Coin Tosses

Consider a common sort of case from the literature:

> COIN TOSS: Suppose Sally could have tossed a fair coin but doesn't. Let p abbreviate 'the coin does not land heads' and let q abbreviate 'the coin does not land tails.' Because Sally does not toss the coin, it does not land heads or tails. Moreover, no one could have ensured that it land heads, nor could anyone have ensured that it land tails. Thus, p is true and no one has a choice in the matter: Np. Likewise for q: thus, Nq. What about the conjunction of p and q? By tossing the coin, Sally could ensure that the coin lands either heads or tails: though true, Sally could render the conjunction of p and q false.

This case, which comes from McKay and Johnson (1996), brings out a piece of common knowledge about the N operator and its logic. Here Np and Nq are both true, but $N(p \wedge q)$ is not. This is therefore a counterexample to the

[6] See also the arguments in Carlson (2000: 284) and Huemer (2000).

[7] This sort of reading might be encouraged by some of Lewis's comments in Lewis (1998). The contrary position we here assume has some support from what van Inwagen says in introducing the N operator in his 1983 book; see van Inwagen (1983: §2.5).

[8] Thanks to an anonymous reviewer for pressing me to clarify this point.

N-theoretic principle of AGGLOMERATION:

$$\text{AGGLOMERATION: } Np, Nq \vdash N(p \wedge q)$$

The central verdict about COIN TOSS is that it presents a counterexample to AGGLOMERATION; thus, AGGLOMERATION is invalid.[9]

But there are other important verdicts at work in the judgment that no one could render either of p or q false. In cases like this, the subject is in no position to do anything, even *given* a toss of the coin, that would have made it the case that the coin lands one way or the other in particular.[10] In the abstract, we say that here Sally can render a conjunction false though she cannot, even having done so, render either of the conjuncts false. Otherwise, in what sense was she *antecedently* unable to render either of the conjuncts false?

The concrete details of Sally's case illustrate this verdict particularly well. Sally cannot render it false that the coin does not land heads, nor can she render it false that the coin does not land tails. Now, suppose Sally tosses the coin. If she is antecedently helpless to bring about a specific outcome, she is no doubt just as helpless to do so *even if* she tosses the coin. Merely tossing the coin is not something that gives her more control over the outcome of a coin toss.

We can regiment this verdict in the following way. We first let Δp abbreviate 'no one can render p false'; as a result, $p \wedge \Delta p$ is equivalent to Np. In COIN TOSS, then, Δp and Δq are both true—this follows from Np and Nq—but they would have *remained* true even if Sally *had* tossed the coin—i.e., if someone had done something to render $(p \wedge q)$ false. This amounts to the following thesis:

$$\text{RENDER} : \forall s \forall \alpha \Big(\big(\text{Does}(s, \alpha) \mathrel{\Box\!\!\rightarrow} \neg(p \wedge q) \big) \rightarrow \big(\text{Does}(s, \alpha) \mathrel{\Box\!\!\rightarrow} (\Delta p \wedge \Delta q) \big) \Big)$$

For anything anyone could have done to render $p \wedge q$ false, had they done so, they would have nonetheless still been unable to render either p or q false. Thus Baker's (2008: 14) characterization of the case: "if [Sally] had tossed [the coin], [Sally] would have no choice about whether it landed heads or tails."

It is not just cases like COIN TOSS that bear out this sort of verdict. Other cases in the literature due, e.g., to Widerker (1987) and Vivhelin (1988) can be easily turned into counterexamples to AGGLOMERATION as well, and these cases of AGGLOMERATION failure furnish further verdicts parallel to RENDER. The phenomenon we are interested in is not parochial—it arises generally in cases that serve as counterexamples to AGGLOMERATION.

RENDER, then, encapsulates an important verdict about the standard counterexamples to AGGLOMERATION. But this is where we find a tension between the counterfactual sufficiency interpretation and widely endorsed logics for counterfactuals. Given the counterfactual sufficiency interpretation and given a moderately strong logic for counterfactuals, RENDER is incompatible with the central

[9] Under specific alternative interpretations of the N operator, this verdict about AGGLOMERATION can be resisted. This will not matter for our purposes, since we are interested in a specific interpretation of the N operator relative to which what I've called the central verdict about COIN TOSS is uncontested.

[10] Statements to this effect can be found in van Inwagen (2000) as well as in Baker (2008).

verdict about the coin-tossing case, *viz.*, that the coin-tossing case is also a case where AGGLOMERATION fails. The proof of incompatibility is laborious, but the central insight behind the proof can be brought out in skeletal form.

5 The Argument: Premises

In COIN TOSS, recall, Np and Nq are true though $N(p \wedge q)$ is false. Given Np and Nq, we are assured that the conjunction $p \wedge q$ is true. By $\neg N(p \wedge q)$ and the counterfactual sufficiency interpretation,[11] then, there is something someone could have done such that, had they done so, the conjunction $p \wedge q$ would have been false. By the description of the case, Sally is one such person, and what she could have done to render $p \wedge q$ false is toss the coin.

Accordingly, letting T abbreviate "Sally tosses the coin", we can give the cash value of $\neg N(p \wedge q)$ in this case in terms of the following conditional:

$$T \mathrel{\Box\!\!\rightarrow} \neg(p \wedge q)$$

If Sally were to toss the coin, then it would not be the case that the coin neither lands heads nor lands tails—it would either land heads or tails. This is our first premise. Given Np, moreover, we are assured that no one, not even Sally, can do anything such that, were she to do it, p would be false. So, in particular, it isn't the case that if she were to toss the coin p would be false:

$$\neg(T \mathrel{\Box\!\!\rightarrow} \neg p)$$

Likewise, by Nq we get the parallel thesis:

$$\neg(T \mathrel{\Box\!\!\rightarrow} \neg q)$$

These are our second and third premises.

That is all by way of unpacking the central verdict—that AGGLOMERATION fails in Sally's coin-tossing case. Our other important verdict, RENDER, tells us that no one would be able to render either p or q false *even if* Sally had tossed the coin. Accordingly:

$$T \mathrel{\Box\!\!\rightarrow} (\Delta p \wedge \Delta q)$$

Given the counterfactual sufficiency interpretation, we can fill out the $\Delta p \wedge \Delta q$ clause occurring in the consequent: $\Delta p \wedge \Delta q$ simply amounts to the pair of theses, first, that nobody (not even Sally) can do anything such that, were they to do it, p would be false, and second, that nobody (not even Sally) can do anything such that, were they to do it, q would be false. If Sally were to toss

[11] Following Pruss (2013), we might capture the sense of Δ, and thereby of N, implied by the counterfactual sufficiency interpretation more thoroughly as

$$\Delta\text{-DEF:} \quad \Delta p =_{df} \neg\exists s \exists \alpha \Big(\mathrm{Can}(s, \alpha) \wedge \big(\mathrm{Does}(s, \alpha) \mathrel{\Box\!\!\rightarrow} \neg p\big)\Big)$$

This definition is adequate to the counterfactual sufficiency interpretation, and an incompatibility proof is available once it is assumed.

the coin, then there wouldn't be anything Sally could do or could have done to render p false; likewise for q. So, in particular, even if Sally had tossed the coin, it isn't the case that, had she tossed the coin, it would have landed heads; likewise for the coin landing tails:

$$T \ \Box\!\!\rightarrow (\neg(T\Box\!\!\rightarrow\neg p\ ) \wedge \neg(T\ \Box\!\!\rightarrow\ \neg q))$$

This is our fourth and final premise.

6 The Argument: Derivation

Everything up to this point has concerned only the verdicts we are interested in, from which our four premises are derived. The incompatibility of our four premises comes out once we accept the following rule governing the counterfactual conditional:

$$\textsc{Centering}^{12}: \ A, \neg(A \ \Box\!\!\rightarrow\ \neg B) \vdash B$$

Centering comes out valid on Lewis's popular logic for counterfactuals. Moreover, it also comes out valid on weaker logics of counterfactuals such as Pollock's (Pollock 1976), and on stronger ones such as Stalnaker's (Stalnaker 1968). If Centering appears unnatural, we note that under Lewis's definition of the so-called *might* counterfactual ($A \Diamond\!\!\rightarrow B =_{df} \neg(A \ \Box\!\!\rightarrow\ \neg B)$), Centering is simply modus ponens for *might* counterfactuals:

$$\textsc{Might-MP}: A, A \Diamond\!\!\rightarrow B \vdash B$$

Lewis's definition of *might* counterfactuals will help us to simplify things in what follows, so we will take Might-MP as a surrogate for Centering.[13]

Given uncontroversial logical assumptions, Might-MP allows us to demonstrate the incompatibility we're after. In particular, we assume two rules valid on all normal logics[14] for conditionals:

$$\textsc{Closure}: \quad \text{if } \ B \vdash C \text{ , then } A \ \Box\!\!\rightarrow\ B \vdash A \ \Box\!\!\rightarrow\ C$$

$$\textsc{Conjunction}: \quad A \ \Box\!\!\rightarrow\ B, \ A \ \Box\!\!\rightarrow\ C \vdash A \ \Box\!\!\rightarrow\ (B \wedge C)$$

We also assume the following axiom, which is common to all but the weakest counterfactual logics of interest:

$$\textsc{Identity}: A \ \Box\!\!\rightarrow\ A$$

[12] Somewhat artificially, Centering is so-called because it is an immediate consequence of Lewis's original centering requirement on sphere models for counterfactuals. See Lewis (1973), pp. 14 for discussion. The weaker requirement called *weak centering* validates the form $A, \neg B \vdash \neg(A \ \Box\!\!\rightarrow\ B)$; the latter delivers us the validity of modus ponens for the counterfactual: $A, A \ \Box\!\!\rightarrow\ B \vdash B$.

[13] In this, we are simply borrowing Lewis's notation; we are not committed to Lewis's idiosyncratic reading of the 'if it were that p, it might be that q' construction.

[14] For an exact characterization of a normal conditional logic, see Chellas (1975).

IDENTITY is useful for establishing the the following helpful (albeit somewhat awkward) form which we call MIGHT-CONTRACTION:

$$\text{MIGHT-CONTR.}: \quad A \boxrightarrow (A \diamondrightarrow B) \vdash A \boxrightarrow B$$

These logical assumptions, taken together, allow us to derive a contradiction from our four premises:

(1)	$T \boxrightarrow \neg(p \wedge q)$	
(2)	$\neg(T \boxrightarrow \neg p)$	
(3)	$\neg(T \boxrightarrow \neg q)$	
(4)	$T \boxrightarrow (\neg(T \boxrightarrow \neg p) \wedge \neg(T \boxrightarrow \neg q))$	
(5)	$T \boxrightarrow \neg(T \boxrightarrow \neg p)$	4; CLOSURE
(6)	$T \boxrightarrow (T \diamondrightarrow p)$	5; Definition
(7)	$T \boxrightarrow p$	6; MIGHT-CONTR.
(8)	$T \boxrightarrow (\neg(p \wedge q) \wedge p)$	1, 7; CONJUNCTION
(9)	$T \boxrightarrow \neg q$	8; CLOSURE
(10)	$(T \boxrightarrow \neg q) \wedge \neg(T \boxrightarrow \neg q)$	3, 9; $\wedge$-Introduction

We made crucial use of the counterfactual sufficiency interpretation in supporting premises 1-4 above. Given CENTERING, these premises turn out to be inconsistent. As promised, then, the counterfactual sufficiency interpretation is in serious tension with CENTERING, a popular thesis about the logic of counterfactuals.

7 Conclusion

The problem is straightforward. We first have the verdict I've called RENDER. Following Baker, we say that Sally would be unable to render p false—and likewise for q—even if she were to toss the coin. This is a verdict about how counterfactuals pattern in cases like Sally's. The counterfactual sufficiency interpretation of the N-operator, however, tells us to interpret 'unable to render false' as itself constraining the patterning of counterfactuals in this case. Given some uncontroversial logical assumptions as well as CENTERING, this in fact lands us in outright contradiction.

We arrived here by assuming the counterfactual sufficiency interpretation of the N-operator. For those unwilling to abandon CENTERING or any of the other more modest assumptions of the argument, it is the counterfactual sufficiency interpretation that has to go. For others, perhaps giving up CENTERING is a more agreeable bargain, though this is a theoretical cost. Either way, what we have here is a problem for the counterfactual sufficiency interpretation.[15]

[15]Many thanks are due to Dominic Lamantia, Fabio Lampert, Paul Manata, Gerard Roth-

University of Notre Dame, USA
jwaldrop@nd.edu

References

Baker, L.R. 2008. "The Irrelevance of the Consequence Argument." *Analysis* 68(1): 13-22.

Cambpell, 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 (2):277-290.

Chellas, B.F. 1974. "Basic Conditional Logic." *Journal of Philosophical Logic* 4(2):133-153.

Cutter, B. 2017. "What is the Consequence Argument an Argument For?" *Analysis* 77(2): 278-287.

Graham, P.A. 2010. "Against the Mind Argument." *Philosophical Studies* 148(2):273-294.

Gustafsson, J.E. 2017. "A Stengthening of the Consequence Argument." *Analysis* 77 (4):705-715.

Finch, A. 2013. "Against Libertarianism." *Philosophical Studies* 166 (3):475-493.

Finch, A., and T.A. Warfield. 1998. "The Mind Argument and Libertarianism." *Mind* 107(427): 515–528.

Huemer, M. 2000. "Van Inwagen's Consequence Argument." *Philosophical Review* 109(4): 525–544.

Lewis, D. 1973. *Counterfactuals*. New York: Blackwell.

Lewis, D. 1981. "Are We Free to Break the Laws?" *Theoria* 47 (3):113-21

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

fus, and three anonymous reviewers for helpful comments on previous drafts of this paper. I am especially indebted to Brian Cutter for insightful comments and much detailed advice during the review process.

Nelkin, D. 2001. "The Consequence Argument and the Mind Argument." *Analysis*, 61 (2):107-115.

Pollock, J.L. 1976. *Subjunctive Reasoning*. Holland: D. Reidel Publishing Company

Pruss, A. 2013. "Incompatibilism Proved." *Canadian Journal of Philosophy*, 43(4): 430-437.

Stalnaker, R.C. 1968. "A Theory of Conditionals." In *Studies in Logical Theory*, edited by Nicholas Rescher. American Philosophical Quarterly Monograph Series 2. Oxford: Blackwell Publishing.

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

Van Inwagen, P. 2000. "Free Will Remains a Mystery." *Philosophical Perspectives* 14, Action and Freedom:1–19.

Van Inwagen, P. 2015. Some Thoughts on An Essay on Free Will. *The Harvard Review of Philosophy* 22: 16-30.

Vihvelin, K. 1988. "The Modal Argument for Incompatibilism." *Philosophical Studies* 53(2): 227–244.

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