A problem not peculiar to counterfactual sufficiency

Chaoan He

Forthcoming in *Analysis*. Please cite published version.

Abstract: The Consequence Argument for incompatibilism is beset by two rival interpretations: the counterfactual sufficiency interpretation and the counterfactual might interpretation. Waldrop recently argued that the counterfactual sufficiency interpretation conflicts with certain principles governing the logic of counterfactuals. In this paper, I show that Waldrop's argument can be adapted to prove that the counterfactual might interpretation also conflicts with the same principles. So the problem Waldrop pointed out is not peculiar to the counterfactual sufficiency interpretation.
Keywords: Consequence Argument, rendering something false, interpretation, counterfactuals

A central debate on the soundness of the Consequence Argument for incompatibilism turns on how to interpret locutions such as '*p* and no one can render *p* false'. There are two main and rival interpretations on offer in the literature (Carlson 2000; Gustafsson 2017; Merlussi 2022): the *counterfactual sufficiency interpretation* and the *counterfactual might interpretation*.

(CSI) An agent *s* can render *p* false if and only if there is something *s* can do such that, were *s* to do it, *p* would be false.

(CMI) An agent *s* can render *p* false if and only if there is something *s* can do such that, were *s* to do it, *p* might be false.

In a recent paper, Waldrop (2023) argued that CSI conflicts with certain widely endorsed principles governing the logic of counterfactuals. In this paper, I show that Waldrop's argument can be adapted to prove that CMI also conflicts with the same principles. So the problem Waldrop pointed out is not *peculiar* to CSI, and the problem by itself does not count as so substantial a challenge to CSI as it initially appears to be.[1]

Waldrop's argument rests on four premises, established via CSI, and a set of principles governing the logic of counterfactuals. Consider the widely discussed case from McKay and Johnson (1996). Suppose Sally could have tossed a fair coin but doesn't. Let p abbreviate 'the coin does not land heads', let q abbreviate 'the coin does not land tails', and let T abbreviate 'Sally tosses the coin'. Since no one could have ensured that the coin lands heads in any given toss, it follows that Sally cannot render p false. This means, according to CSI, that there is nothing Sally could do such that, were she to do it, p would be false. In particular, it is not the case that if she were to toss the coin, p would be false. Likewise for q. Thus, the following two propositions obtain:

(Premise A) $\neg(T \Box\!\!\rightarrow \neg p)$
(Premise B) $\neg(T \Box\!\!\rightarrow \neg q)$

But since the coin would either land heads or land tails in any given toss, Sally can toss the coin and ensure that it either lands heads or lands tails. That means Sally can render $p \wedge q$ false. According to CSI, then, there is something Sally could do such that, were she to do it, $p \wedge q$ would be false. In particular, if Sally were to toss the coin, $p \wedge q$ would be false. Or:

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

Finally, not only is Sally unable to render p or q false, but her inability (to render p or q false) would have *remained* unchanged even if Sally *had* tossed the coin. For abilities (and inabilities)

[1] Waldrop's paper and the present paper are only concerned with the left-to-right direction of both CSI and CMI. The right-to-left direction has been criticized, by Schnieder (2004), Hausmann (2018) and De Rizzo (2002), among others, for considerations having to do with relevance and agency. For instance, the right-to-left direction implies that I can render 2+2=5 false, for there is something I can do, raising my left arm, say, such that, if I did it, 2+2=5 would/might be false.

are relatively stable qualities had by agents, not easily changeable. Merely tossing the coin simply is not something that would change Sally's inability and give her more control over the outcome of a toss. That means, according to CSI, even if Sally had tossed the coin, there would not be anything Sally could do such that, were she to do it, *p* would be false, and there would not be anything Sally could do such that, were she to do it, *q* would be false. In particular, even if Sally had tossed the coin, it is not the case that, had she tossed the coin, *p* would have been false, and it is not the case that, had she tossed the coin, *q* would have been false. This establishes the final premise for Waldrop's argument.

(Premise D) $T \;\Box\!\!\rightarrow (\neg(T \Box\!\!\rightarrow \neg p) \wedge \neg(T \Box\!\!\rightarrow \neg q))$[2]

Waldrop's argument also rests on a set of principles governing the logic of counterfactuals:

(Definition) $A \Diamond\!\!\rightarrow B =_{df} \neg(A \Box\!\!\rightarrow \neg B)$

(Closure) if $B \vdash C$, then $A \;\Box\!\!\rightarrow B \vdash A \;\Box\!\!\rightarrow C$

(Conjunction) $A \;\Box\!\!\rightarrow B, A \;\Box\!\!\rightarrow C \vdash A \;\Box\!\!\rightarrow (B \wedge C)$

[2] From the point of view of modal semantics, Premise D may seem much less plausible than Premises A-C. Especially, to anyone who assumes that even the match of particular facts (such as a coin's landing heads in a certain toss) matter in measuring the similarity between possible worlds, D may appear very implausible. This is because, either $T \Box\!\!\rightarrow (T \Box\!\!\rightarrow \neg p)$ or $T \Box\!\!\rightarrow (T \Box\!\!\rightarrow \neg q)$ has to be true under that assumption. To see this, consider any world *w* where Sally tossed the coin and it landed heads. If the fact that the coin landed heads is held fixed and counts in measuring worlds-similarity, then all the closest worlds to *w* where *T* obtains are such that the coin lands heads. But that plainly means that $T \Box\!\!\rightarrow (T \Box\!\!\rightarrow \neg p)$ is true, which contradicts D. One way to get around this problem is by simply embracing the idea that the match of certain particular facts does not matter, at least in some cases, in measuring worlds-similarity. Another way, which is perhaps less controversial, is to adopt some kind of past tracking reading of D, to such an effect that when evaluating the embedded conditional $T \Box\!\!\rightarrow \neg p$, we go back to the point where Sally still had *not* tossed the coin. Then the set of closest worlds to *w* where *T* obtains includes both worlds where the coin lands heads and worlds where it lands tails, which would secure Premise D. So in view of such considerations, one may still be inclined to continue working with Premise D and see where it leads, as Waldrop does.

(Might-contraction) $A \Box\!\!\rightarrow (A \Diamond\!\!\rightarrow B) \vdash A \Box\!\!\rightarrow B$

Most notably, Might-contraction itself rests on the principle of Centering:

(Centering) $A, \neg(A \Box\!\!\rightarrow \neg B) \vdash B$ [3]

Now, a contradiction logically follows from the afore-mentioned premises and principles:

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

(2) $\neg(T \Box\!\!\rightarrow \neg p)$ Premise A

(3) $\neg(T \Box\!\!\rightarrow \neg q)$ Premise B

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

(5) $T \Box\!\!\rightarrow \neg(T \Box\!\!\rightarrow \neg p)$ 4; Closure

(6) $T \Box\!\!\rightarrow (T \Diamond\!\!\rightarrow p)$ 5; Definition

(7) $T \Box\!\!\rightarrow p$ 6; Might-contraction

(8) $T \Box\!\!\rightarrow (\neg(p \wedge q) \wedge p)$ 1,7; Conjunction

(9) $T \Box\!\!\rightarrow \neg q$ 8; Closure

(10) $(T \Box\!\!\rightarrow \neg q) \wedge \neg(T \Box\!\!\rightarrow \neg q)$ 3,9; $\wedge$-Introduction

Waldrop takes the argument to be a straightforward problem for CSI. As the premises are established via CSI, if we are unwilling to abandon the principles governing the logic of counterfactuals, CSI would appear to prove false. In what follows, we will see that Waldrop's

[3] A more common statement of Centering is that any world is more similar to itself than any other world is to it. Though highly controversial in the literature, Centering can be tentatively justified on a number of grounds. For one thing, as Bennett (1974) and his many followers claim, Centering is very plausible at an intuitive level, since no world can be as similar to a given world *w* as *w* itself is. For another, as Walters (2016) and McDermott (2007) rightly observe, there are indisputable cases of counterfactuals that seem to require Centering. Still, on the standard Lewis-Stalnaker semantics, Centering validates the principle of Conjunction Conditionalization, which, as Walters and Williams (2013) argue, follows from a certain package of theorems of the standard logic of counterfactuals such as Lewis' system VW.

argument can be adapted to show that CMI, the chief rival of CSI, also conflicts with the same principles.

Waldrop's argument appeals to some intuitive verdicts concerning the McKay & Johnson case. For instance, it is intuitively true that Sally cannot render either *p* or *q* false, in view of the fact that she can neither ensure that the coin *would* land heads nor ensure that it *would* land tails in any given toss. Such intuitive verdicts, one should note, accord well with the CSI interpretation of 'can render something false'. The McKay & Johnson case was originally developed to undermine Van Inwagen's argument for incompatibilism, construed along CSI lines. An influential incompatibilist reaction to the McKay & Johnson case is to adopt the CMI interpretation of 'can render something false' and construe Inwagen's argument accordingly. (Gustafsson 2017; Merlussi 2022) On this alternative approach, which many philosophers (including Van Inwagen himself) adopted, Sally can render both *p* and *q* false, for she is able to act so that *p might* be false and *q might* be false.

This is not to question the purported truth *per se* of the intuitive verdicts Waldrop's argument appealed to. On the contrary, I think Waldrop is perfectly entitled to appeal to the intuition (that Sally cannot render either *p* or *q* false) when attempting to pose a problem for CSI, for the intuition accords well with CSI and appears hardly disputable for advocates of CSI. By the same token, however, we may equally help ourselves to the contrary intuitive verdicts (that Sally can render both *p* and *q* false) when we attempt here to pose a problem for CMI, given that those intuitions accord well with CMI and are generally embraced by advocates of CMI. The contrary intuitive verdicts, interpreted according to CMI, imply that there is something Sally could do, such that, were she to do it, *p* might be false and *q* might be false. In particular, if she were to toss the coin, *p* might be false and *q* might be false. So in place of Waldrop's Premise A and Premise B, we shall adopt the following pair of premises:

(Premise A*) $T \Diamond\!\!\rightarrow \neg p$

(Premise B*) $T \Diamond\!\!\rightarrow \neg q$

Now consider a third premise for our argument. Given our supposition that Sally did not toss the coin, p and q are both true, hence $p \vee q$ is also true. So is Sally able to render $p \vee q$ false? If yes, then according to CMI, Sally is able to do something such that, were she to do it, $p \vee q$ might be false. Since for $p \vee q$ to be false is for the coin to land heads *and* tails, which is impossible to obtain, it follows that Sally cannot render $p \vee q$ false. So it is not the case that there is something Sally can do such that, were she to do it, $p \vee q$ might be false. In particular, it is not the case that if Sally were to toss the coin, $p \vee q$ might be false.

(Premise C*) $\neg(T \Diamond\!\!\rightarrow \neg(p \vee q))$

Our last premise is justifiable in a way Waldrop's Premise D is justified. Since Sally is able to render both p and q false, her ability (to render p and q false) would have *remained* unchanged even if Sally had tossed the coin. Merely tossing the coin is not something that would impair Sally's ability in question. This means, according to CMI, even if Sally had tossed the coin, there would be something Sally could do such that, were she to do it, p might be false, and there would be something Sally could do such that, were she to do it, q might be false. In particular, even if Sally had tossed the coin, it is the case that, had she tossed the coin, p might have been false, and had she tossed the coin, q might have been false. Thus, the following obtains.

(Premise D*) $T \Box\!\!\rightarrow ((T \Diamond\!\!\rightarrow \neg p) \wedge (T \Diamond\!\!\rightarrow \neg q))$[4]

[4] Premise D* may be disputed on the same grounds as Premise D is disputed. But to the extend that the two premises are on a par in relevant respects, the argument to be presented below, if other things being equal, would be no less plausible than Waldrop's original argument.

Now, assuming the principles of Definition, Closure, Conjunction and Might-contraction, as Waldrop's argument does, a contradiction can be derived from the four premises:

(1) $\neg(T \Diamond\!\!\rightarrow \neg(p \vee q))$	Premise C*
(2) $T \Diamond\!\!\rightarrow \neg p$	Premise A*
(3) $T \Diamond\!\!\rightarrow \neg q$	Premise B*
(4) $T \Box\!\!\rightarrow ((T \Diamond\!\!\rightarrow \neg p) \wedge (T \Diamond\!\!\rightarrow \neg q))$	Premise D*
(5) $T \Box\!\!\rightarrow (T \Diamond\!\!\rightarrow \neg p)$	4; Closure
(6) $T \Box\!\!\rightarrow \neg p$	5; Might-contraction
(7) $T \Box\!\!\rightarrow (p \vee q)$	1, Definition
(8) $T \Box\!\!\rightarrow (\neg p \wedge (p \vee q))$	6,7; Conjunction
(9) $T \Box\!\!\rightarrow q$	8; Closure
(10) $\neg(T \Box\!\!\rightarrow q)$	3, Definition
(11) $(T \Box\!\!\rightarrow q) \wedge \neg(T \Box\!\!\rightarrow q)$	9; 10; $\wedge$-Introduction

To conclude, Waldrop's argument purports to show that CSI conflicts with a set of principles governing the logic of counterfactuals. I developed an argument here, in the image of Waldrop's, which shows that CMI, the chief rival of CSI, also appears to conflict with the same principles. So the problem Waldrop pointed out is not a problem *peculiar* to CSI. It is a *general* problem, if at all, for friends and foes of CSI alike.[5]

Donghua University
China
chaoanhe@hotmail.com

[5] I would like to thank two anonymous referees of this journal for helpful comments on an earlier version of this paper. Versions of this paper were presented at workshops in Shandong University (Qingdao) and Tongji University. Thanks to all the participants for valuable discussions.

References

Bennett, J. 1974. Counterfactuals and possible worlds. *Canadian Journal of Philosophy* 4: 381-402.

Carlson, E. 2000. Incompatibilism and the transfer of power necessity. *Noûs* 34: 277-90.

De Rizzo, J.2022. No choice for incompatibilism. *Thought: A Journal of Philosophy* 11: 6-13.

Gustafsson, J. 2017. A strengthening of the consequence argument for incompatibilism. *Analysis* 77: 705-15.

Hausmann, M. 2018. The consequence argument ungrounded. *Synthese* 195: 4931-50.

McDermott, M. 2007. True antecedents. *Acta Analytica* 22: 333-35.

McKay, T. and D. Johnson. 1996. A reconsideration of an argument against compatibilism. *Philosophical Topics* 24: 113-22.

Merlussi, P. 2022. Revisiting McKay and Johnson's counterexample to beta. *Philosophical Explorations* 25: 189-203.

Schnieder, B. 2004. Compatibilism and the notion of rendering something false. *Philosophical Studies* 117: 409-28.

Waldrop, J. 2023. A problem for counterfactual sufficiency. *Analysis* 83: 527-35.

Walters, L. 2016. Possible worlds semantics and true-true counterfactuals. *Pacific Philosophical Quarterly* 97: 322-46.

Walters, L. and R. Williams. 2013. An argument for conjunction conditionalization. *Review of Symbolic Logic* 6: 573-88.