Newcomb's Problem: A Reply to Carlson
Author(s): John Martin Fischer
Source: *Analysis*, Vol. 61, No. 3 (Jul., 2001), pp. 229-236
Published by: Oxford University Press on behalf of The Analysis Committee
Stable URL: http://www.jstor.org/stable/3329238
Accessed: 24/02/2011 13:38

> we have no sufficient grounds for saying that [claims similar to a certain claim which, like (1) and (2), contains 'actual'] express such propositions. (2000: 192, n. 6)

No similar retort to the accusation that Stephanou's argument for the necessity of (1) and (2) fails can work, however. For, as §3 showed, (1) and (2) are not both necessary, which certainly gives us sufficient grounds for not inferring their necessity from the fact that they are logical or conceptual truths. Stephanou's inference to the necessity of (1) and (2) is an overreaction to their special logical or conceptual properties.

The upshot of this discussion is a happy one. It would be startling to find that one can regard Aristotle's existence as contingent only if one thinks that there are possible necessities which are untrue. Thankfully, however, we can remain unshaken. Stephanou's argument (Y) is indeed unsound, but not for the reasons which he gives.[1]

Trinity College
Cambridge CB2 1TQ, UK
dig21@hermes.cam.ac.uk

Reference

Stephanou, Y. 2000. Necessary beings. *Analysis* 60: 188–93.

[1] Thanks to Michael Clark, Rosanna Keefe, Alex Oliver, Yannis Stephanou and an anonymous referee for helpful discussion.

Newcomb's problem: a reply to Carlson

JOHN MARTIN FISCHER

It is often thought that the correct solution to Newcomb's problem depends on whether or not one should resolve the vagueness of counterfactual conditionals in the context of this puzzle in such a way as to vindicate backtracking. If the relevant counterfactuals backtrack, then it seems rational to choose the one box. But if they do not backtrack (and thus the predictor's prediction is not counterfactually dependent on my choice), then the two-box approach seems to be correct.

In *The Metaphysics of Free Will* I proposed a strategy according to which we can side-step what can seem to be a stalemate concerning the conditionals in question (Fischer 1994: 98–107). I claimed that on this approach

ANALYSIS 61.3, July 2001, pp. 229–36. © John Martin Fischer

one can defend the two-box solution quite independently of any assumptions about the counterfactuals (given an inerrant predictor). Further, I claimed that this way of thinking about the puzzle issues in an asymmetry between the context of a merely inerrant (actually always correct) predictor and a genuinely infallible (necessarily always correct) predictor; if the predictor is merely inerrant, the two-box solution is rational, but if the predictor is infallible, then the one-box solution is rational.

Erik Carlson takes issue with my argument that one can defend the two-box strategy even while granting the truth of the relevant backtracking counterfactuals.[1] Carlson says that he will 'try to show that [Fischer's] argument, although ingenious, does not succeed' (Carlson 1998: 229). In this brief piece I aim to show that (only!) the latter part of Carlson's claim is false.

My approach to defending the rationality of the two-box solution in the context of an inerrant (but not infallible) predictor employs the assumptions that an agent can perform only those actions that are extensions of the actual past, and thus that the only reasons relevant to an individual's deliberations are those which obtain in the worlds that are extensions of the actual past. Given this extremely plausible constraint on our abilities, and the associated connection with practical reasoning and the relevance of reasons, I presented the following simple argument for the two-box solution. Either the predictor put the one million dollars – $M – in Box (B2) or not. If he did, then the only possible worlds I can now actualize have in their past that this is so, and I get $1,000 more in the one of these in which I take both boxes. If he did not, then the only possible worlds I can now actualize (and thus relevant to my practical reasoning) have in their past that this is so, and I at least get $1,000 in the one of these in which I take both boxes. Thus, although I do not know this feature of the actual past, there is a dominance strategy in the sense that, on *either* assumption about the past, I should take the two boxes. And this is so, whatever may be the case with respect to the counterfactuals. Rather than making use of the counterfactuals, my approach simply looks at the reasons that are present in accessible possible worlds (i.e. those worlds which – among other things – share the same past as the actual world).

[1] Whereas Carlson focuses on the defence of the two-box approach in the context of inerrancy, Jordan Howard Sobel criticizes the asymmetry thesis (and, in particular, my contention that my approach entails the rationality of the one-box solution in the context of infallibility) in Sobel 1998a: 96–100; and 1998b: 172. I give a reply to Sobel's criticism, and a general defence of my approach to Newcomb's Problem (including the context in which the predictor is assumed to be *infallible*) in (Fischer forthcoming).

Now Carlson argues against my approach as follows (1998: 230). Suppose there is in fact counterfactual dependence of the predictor's predictions and my choices, and thus that the backtrackers are true:

(1) If I were to choose the second box, then the predictor would have put $M in the second box.
(2) If I were to choose both boxes, then the predictor would have left the second box empty.

Given my assumptions about accessibility and reasons, however, the following holds at the time when I am to make my choice:

(3) Either no world where the predictor put $M in the second box is accessible to me, or no world where the predictor left the second box empty is accessible to me.

Carlson now points out that I contend that the conjunction of the above three claims is compatible with the conjunction of the following 'can-claims':

(4) I can choose the second box.
(5) I can choose both boxes.

But Carlson disagrees, claiming that (3) is incompatible with the conjunction of (1), (2), (4) and (5). This is his argument:

> Intuitively, it seems clear that if I now have an option which is such that if I were to choose it a certain world would be actual, then this world is now accessible to me. [In his n. 3 Carlson claims that this follows from Fred Feldman's 'reflexivity' and 'transitivity' axioms of accessibility, in Feldman 1986: 20–1; I shall return to this claim below.] All I have to do to actualize it is to choose this particular option. Then, surely, I *can* actualize this world. If we accept this claim it follows from (1), (2), (4), and (5) that I have access both to a world where the predictor put $M in the second box, and to a world where he left it empty, and hence that (3) is false. (1998: 230)

The crucial claim here is, 'if I now have an option which is such that if I were to choose it a certain world would be actual, then this world is now accessible to me. All I have to do to actualize it is to choose this particular option'. But I do not believe that the conditional, 'if I were to choose a given option, then a certain world would be actual' entails the accessibility claim, 'I now have access to the possible world in question (mentioned in the consequent of the conditional).' Consider, for example, a context in which I have a genuine inability to choose to pick up a snake, owing to a traumatic event involving a snake in my youth. Given this past event, I just cannot bring myself to choose to pick up the snake. Now this is just the sort of context in which it would seem that the appropriate resolution of the vagueness of counterfactuals would lead to the vindication of the backtracker: if I were to choose to pick up the snake now, this would show that the traumatic event never happened in the past. That is, if I were to pick up the snake, it would be in virtue of lacking the pathological aversion – and this would show that the traumatic event never happened. But none of these truths about counterfactuals entails that a world in which I choose to pick up the snake (and do so) is now *accessible* to me.

It may well be that, if I were to choose to do something, that would show the lack of some obstacle to choosing it that actually exists; obviously it does *not* follow that I actually can choose the thing in question (or have access to a world in which I so choose); after all, the obstacle *actually exists*. Similarly, the mere fact that, if I were to choose something, that would show that the past was different from what it actually was does *not* entail that I *can* choose the thing in question or that I have access to a possible world in which I so choose; the supposition (of my argument in *The Metaphysics of Free Will*) is that I can only 'get to' possible worlds which are extensions of the actual past.

To motivate further the point I am making about accessibility, note that (regrettably) John F. Kennedy was assassinated in 1963. My contention is that any possible world I can now actualize must be an extension of this world, and thus must have in its past that John F. Kennedy was assassinated. If 'the future is a garden of forking paths', these paths branch off a *single* past – that, at least, is the intuitive picture of the structure of agency and possibility over time. Any world which does not have in its past that John F. Kennedy was assassinated in 1963 is a world which is such that *I can't get there from here*, so to speak. So, if I really can choose to pick up the snake (in the example above), then my choosing to pick up the snake must be an extension of a world in which the traumatic event happened in my past. The mere fact that, if I were to choose to pick up the snake, I would not have had that traumatic experience does *not* entail that my choosing to pick up the snake can be an extension of the *actual* world in which *I did have that experience*.

Carlson also offers what he calls a 'related but less direct argument' against the compatibility of (1) through (5):

> Trivially, any world is accessible from itself. This means that if I would actualize a certain world, this world would be accessible to me. Hence, (1) and (2), respectively, yield
>
> (1*) If I were to choose the second box, then a world where the predictor put $M in the second box is accessible to me, and
> (2*) If I were to choose both boxes, then a world where the predictor left the second box empty is accessible to me.

> The only way to reconcile (3) with the conjunction of (1*), (2*), (4), and (5) is to maintain that what worlds are accessible to me at *t* depends on which of my options I do in fact choose at *t*. Fischer must claim that a world where the predictor put $M in the second box would be accessible to me only if I were to choose the second box, whereas a world where the predictor left the second box empty would be accessible only if I were to choose both boxes. To hold that the accessibility of certain worlds is counterfactually dependent on my choice seems very implausible. (1998: 230–31)

Carlson's argument is essentially as follows. Assume that I can choose the second box, and that I can choose both boxes. Given (1*), then, I can perform an action which is such that if I were to perform it, I would have access to a world in which the predictor put $M in the second box. And given (2*), I can perform an action which is such that, if I were to perform it, I would have access to a world in which the second box is empty. So (3) cannot be true.

But given the assumption with which we are working here (that the predictor is inerrant but not infallible), I deny that I must hold the view Carlson attributes to me. Just as Carlson's second argument is 'related' to his first, there is a related reply. Note that the key move in Carlson's argument is from 'I can perform an action *X* such that were I to perform it, I would have access to a certain possible world *W*' to 'I have access to *W*.' But this is exactly the transition that I am at pains to deny. Suppose I can perform *X*. On my approach, it follows that there is a possible world with the same past as the actual world in which I do *X*; I have access to this 'same-past' world *W** in so far as it is an extension of the actual past. Now the counterfactual, 'If I were to do *X*, I would have access to some world *W* in which the past is different from the actual past', may be true *without* thereby entailing that I have access to *W*.

The point of my argument in *The Metaphysics of Free Will* is that the possible world(s) relevant to the counterfactual need not be the same as those relevant to the 'can-claim' and therefore the facts about accessibility. The possible worlds relevant to the counterfactual (in virtue of which the counterfactual is true, if it is true) are the 'closest' or 'most similar' to the actual world in which I do *X*; the possible worlds relevant to the can-claim, and thus to which the agent has the pertinent sort of access, *need not be the closest worlds in which I do X*. Rather, they *do* need to be extensions of the actual past.[2] Given this set of claims about the logical terrain, Carlson's contention that it follows from 'I can perform an action *X* such

[2] Of course, I cannot here give the full development of this argument; for the argument, see Fischer 1994: 87–110.

that, were I to perform it, I would have access to a certain possible world *W*' that 'I have access to *W*' is false.

In the example of the snake above we noted that, if I were to choose to pick up the snake, some actually existing obstacle to my choosing to pick up the snake would be gone, and some actually occurring past event *would not have taken place*. But (as I pointed out above) this does not show that I actually have access to a possible world in which some actually occurring past event would not have taken place. The consequent of the counterfactual specifies a condition obtaining in the closest possible world(s) in which I pick up the snake; but these need not be the worlds to which I have access. Similarly, the consequent of the conditional, 'If I were to do *X*, I would have access to a possible world with a different past from the actual past', specifies a condition that obtains in the closest possible world(s) in which I do *X*; but these need not be the worlds to which I *actually do have access*. Recall that Carlson says, 'Trivially, any world is accessible from itself. This means that if I would actualize a certain world, this world would be accessible to me'. Fine; but one needs to distinguish between what *is* accessible to me and what *would be* accessible (under certain counterfactual suppositions). Carlson's argument shows that certain worlds *would be* accessible under certain counterfactual suppositions; but the thrust of my argument shows that it does not follow that those worlds *are* accessible.

Return to Carlson's crucial claim of his first argument: 'if I now have an option which is such that if I were to choose it a certain world would be actual, then this world is now accessible to me. All I have to do to actualize it is to choose this particular option.' As we saw above, we have it on Carlson's authority that this follows from certain axioms presented by Fred Feldman. Surely, if it is plausible that this claim follows from Feldman's axioms, we must insert the assumption that I *can* take the option in question (make the choice, perform the action, and so forth). But in fact the claim (even suitably revised) does *not* follow from these axioms, as far as I can see. This is because Feldman's axioms pertain to combinations of propositions about accessibility, whereas Carlson's claim pertains to the relationship between propositions about accessibility and certain counterfactuals. But, as I argued above, the consequent of the relevant counterfactual *need not* point us to an accessible possible world.

To see why Feldman's axioms are of no help to Carlson, consider this quotation from Feldman, which helps to explain the import of his axiom of transitivity:

> It is important to recognize that [the transitivity axiom] does not imply, for example, that if *s* can work in his garden, and if he were to work in his garden, then he would be able to harvest his crop, that he

> can already harvest his crop. What [the axiom of transitivity] does mean is that if a world in which he gardens is accessible to *s* from here now, and if a world in which he harvests is accessible to him from there now, then the world in which he harvests is also accessible to him from here now. (1986: 21)

What is important to me is not the particular contrast Feldman here draws, which (in part) is about complications issuing from temporal considerations, but the point that Feldman is talking about the structure of combinations of propositions about *accessibility*. So, on Feldman's view, if world *W1* is accessible to me now, and *W2* is accessible to me now from *W1*, then *W2* is accessible to me now. I can accept this *without* thereby accepting Carlson's problematic claim (even suitably revised as suggested above): if *W1* – a world in which I perform some act – is accessible to me now, and if it is true that if I were to perform that act, then *W2* would be actual, then it follows that I now have access to *W2*. As I have emphasized above, the worlds in virtue of which the counterfactual is true are the *closest* possible worlds (to the actual world) in which I perform the act in question; this set of worlds need not include a world with the same past as the actual world, and thus need not include a possible world accessible to me from the actual world. So even though *W1* is accessible to me, it need not be the case that *W2* is accessible to me from *W1*, and thus it does *not* follow that *W2* is accessible to me from the actual world. Feldman's axioms therefore do not help to establish the crucial contention of Carlson's argument.

Thus, with respect to a merely inerrant predictor, I need *not* claim that a possible world's accessibility depends on what the agent chooses. Rather, it *does* depend on the past in the sense that a possible world's accessibility requires that it be an extension of the actual past. Carlson twice calls my conception of accessibility 'peculiar'. For example, he says, 'On this peculiar conception, which kind of world is accessible to me depends on what I choose to do' (1998: 231). I have shown that my approach does not entail that which world is accessible to me depends on what I choose (given the assumption of the predictor's mere inerrancy). If *this* is what Carlson finds peculiar, then there is no objection to my approach. Further, *there is absolutely nothing peculiar* about the contention that the only possible scenarios relevant to an agent's freedom at a given time are extensions of the past relative to that time. Given the intuitive picture of the branching structure of human agency and possibilities, what would be peculiar is its *denial*.

University of California
Riverside, California 92521, USA
John.Fischer@ucr.edu

References

Carlson, E. 1998. Fischer on backtracking and Newcomb's problem. *Analysis*. 58: 229–31.

Feldman, F. 1986. *Doing the Best We Can*. Dordrecht: D. Reidel Publishing Company.

Fischer, J. M. 1994. *The Metaphysics of Free Will: An Essay on Control*. Oxford: Blackwell.

Fischer, J. M. Forthcoming. Critical notice of J. H. Sobel, *Puzzles for the Will*. *Canadian Journal of Philosophy*.

Sobel, J. H. 1998a. Critical notice of John Martin Fischer, *The Metaphysics of Free Will*. *Canadian Journal of Philosophy* 28: 95–117.

Sobel, J. H. 1998b. *Puzzles for the Will*. Toronto: University of Toronto Press.

Why coherence is not truth-conducive

ERIK J. OLSSON

Tomoji Shogenji (1999) argues that beliefs that are more coherent need not thereby be more likely to be true. This would strongly suggest that coherence (in the sense of mutual support) is not truth-conducive, that is, that the traditional philosophical problem of whether coherence implies truth has been given a negative solution. But Shogenji resists this conclusion, maintaining that the notion of truth-conduciveness is inadequately captured by the formula 'more coherence implies a higher likelihood of truth'. Further, he goes on to show that coherence *is* truth-conducive on another, supposedly more adequate, account of the problematic concept. In this paper I shall argue that, while Shogenji's observation of a lack of correlation between coherence and truth is correct, his reasons for not drawing the natural philosophical conclusion are not.

Shogenji has drawn attention to the following problem of coherence and specificity. He asks us to imagine 'an epistemically ultraconservative agent who only holds a few extremely unspecific beliefs – say, some rocks are heavier than others; some animals sleep sometimes; and someone is humming some tune somewhere'. In this case it is, Shogenji writes, 'very likely that her beliefs are all true even though they do not hang together'. Meanwhile 'a huge collection of highly specific beliefs – such as the entire body of medical science – almost certainly contains errors even though they tightly hang together' (1999: 342). Hence, more coherence does not imply a higher joint likelihood of truth.

ANALYSIS 61.3, July 2001, pp. 236–41. © Erik J. Olsson