Prompt: *Explain van Inwagen's "consequence argument". Describe what you think is the best response to this argument. Does this response succeed in saving compatibilism from the consequence argument? Why or why not?*

A "Can" of Worms

by William Hoza

In "The Incompatibility of Free Will and Determinism" [1], Peter van Inwagen argues that *if* the universe is deterministic, *then* free will does not exist. (He is silent about whether the universe is in fact deterministic and about whether free will in fact exists.) This is in contrast to the *compatibilist* position, which holds that free will and determinism are not contradictory. Briefly, van Inwagen's argument is that when an agent with free will performs some action, she (by definition of "free will") could have performed a different action. But in a deterministic universe, acting a different way requires either altering the past or violating the laws of physics. So van Inwagen concludes that the free agent could have either altered the past or violated the laws of physics. Finally, van Inwagen says that it is obvious that nobody can alter the past, and by definition of the phrase "law of physics", nobody can violate the laws of physics either. So our hypothetical free agent in a deterministic universe cannot exist.

I will argue that van Inwagen's argument does not succeed in ruling out the possibility of free will in a deterministic universe, because his idea of what it means to say that an agent "could have done otherwise" is not the relevant notion for free will. When the phrase "could have" is analyzed correctly (using the standard "conditional analysis"), van Inwagen's argument breaks down in a perhaps counterintuitive way: From the facts that an agent could have done X and that doing X requires doing Y, it does not follow that the agent could have done Y.

I'll begin by explaining van Inwagen's argument in more detail. He defines *determinism* to be "the conjunction of these two theses: (a) For every instant of time, there is a proposition that expresses the state of the world at that instant. (b) If A and B are any propositions that express the state of the world at some instants, then the conjunction of A with the laws of physics entails B." [2, page 214] He leaves the phrase "law of physics" undefined, but promises to only make assumptions about the phrase which any reasonable person would agree must come out true under any legitimate analysis of the phrase. I have no complaints with van Inwagen's usage of the phrase.

Next, van Inwagen discusses what it means to have free will. He says that at the least, if an agent performs an act freely, then "that agent *could have* refrained from performing the act." [2, page 216] As with the phrase "law of physics", van Inwagen declines to give a definition for the relevant sense of "can"/"could have".[1] Again, he promises to only make uncontroversial assumptions: [2, page 217]

> There is, however, considerably less agreement as to how 'can' (in the relevant sense) should be analysed. This is one of the most difficult questions in philosophy. It is certainly a question to which I do not know any nontrivial answer. But, as I said I should do in the case of 'law of physics', I shall make certain conceptual claims about 'can' (in the 'power' or 'ability' sense) in the absence of any analysis. Any suggested analysis of 'can' that does not support these claims will either be neutral with respect to them, in which case it will be incomplete, since it will not settle all conceptual questions about 'can', or it will be inconsistent with them, in which case the arguments I shall present in support of these claims will, in effect, be arguments that the analysis fails.

Later, I will indeed argue in favor of an analysis of "can" that does not support van Inwagen's claims about "can".

Finally, van Inwagen presents his argument. He imagines, as a prototypical example of a free-seeming action, a judge J who decides at time T to not raise his hand. Following van Inwagen, let T_0 be a time prior to J's birth, let P and P_0 be propositions expressing the states of the universe at times T and T_0 respectively, and let L be the conjunction of all the laws of physics. Van Inwagen then argues as follows that if determinism is true, then J could not have raised his hand at time T: Assume for a contradiction that determinism is true and J could have raised his hand at time T. By determinism, $(P_0 \wedge L)$ entails P, and hence (by contraposition) $\neg P$ entails $(\neg P_0 \vee \neg L)$. Since J could have raised his hand at time T, he could have made $\neg P$ true. Then – and this is the crucial step which I will criticize – van Inwagen says, from the facts that J could have made $\neg P$ true and that $\neg P$ entails $(\neg P_0 \vee \neg L)$, we may conclude that J could have made $(\neg P_0 \vee \neg L)$ true. But, van Inwagen says, that's absurd; J can neither affect things that took place before his birth, nor violate the laws of physics.

Van Inwagen discusses most steps of this argument in great detail, but he only briefly defends the step with which I take issue: [2, page 221]

> This premiss [that if J could have rendered P false, and if $(P_0 \wedge L)$ entails P, then J could have rendered $(P_0 \wedge L)$ false] may be defended as an instance of the following general principle:
>
> > If S can render R false, and if Q entails R, then S can render Q false.
>
> This principle seems to be analytic. For if Q entails R, then the denial of R entails the denial of Q. Thus, any condition sufficient for the falsity of R is also sufficient for the falsity of Q. Therefore, if there is some condition that S can produce that is sufficient for the falsity of R, there is some condition (that same condition) that S can produce that is sufficient for the falsity of Q.

[1] Following van Inwagen, I will treat the phrases "can" and "could have" as differing only grammatically.

I'll refer to this "general principle" – the principle that the set of statements that an agent can render true is closed under entailment – as "van Inwagen's entailment principle". Van Inwagen's entailment principle sounds plausible at first, but van Inwagen's defense of the principle is circular. To try to establish that S can render Q false, he merely establishes that there is some condition that S can produce that is *sufficient* for the falsity of Q. Presumably, the sense of sufficiency meant here is that the condition logically entails the falsity of Q. Thus, to arrive at the desired conclusion that S can render Q false, van Inwagen would need precisely the entailment principle that he is trying to prove, with R being the negation of the aforementioned sufficient condition.

I think the real reason that van Inwagen's entailment principle is tempting is because it is true for a certain notion of "can". Sometimes, when a person says that something *can* happen, she merely means that it is *logically possible* for the thing to happen. (E.g. "The Earth could have been invaded by aliens in 1941, in which case World War II would have gone much differently.") With respect to this notion of "can", van Inwagen's entailment principle is certainly true. But this notion of "can" is not the relevant one for free will. The fact that Mount St. Helens could have erupted today does not help Mount St. Helens to have free will.

In fact, I will argue that van Inwagen's entailment principle is *false* when "can" is interpreted in the sense relevant to free will. And what sense is that? In the Mount St. Helens example, the missing ingredient for free will is that even in our hypothetical scenario where Mount St. Helens erupted today, *it still wasn't on purpose.* The sense of "can" relevant to free will is an *intention*-oriented sense. To be more specific, I, like countless compatibilists before me, advocate the so-called "conditional analysis" of "can": To say that an agent can do X means that if the agent formed the intention to do X, then she would do X.

Notice that when an agent forms an intention to do X, even if X entails Y, the agent might not automatically form an intention to do Y. For example, you might form an intention to write down a prime p such that $p \equiv 1 \pmod 4$ without forming the intention to write down a prime which is a sum of two squares, despite a theorem of Fermat that states that the two conditions are equivalent.

This observation about intentions hopefully makes van Inwagen's entailment principle sound much less plausible. In particular, the instance of van Inwagen's entailment principle that appears in his main argument should now sound dubious and in serious need of justification. Clearly, the judge truly can raise his hand: if he were to form the intention to raise his hand, he would succeed (*because* of the laws of physics!) Can the judge alter circumstances before his birth? That depends on your model of counterfactual conditionals. But the point is that the relevant question – "What would happen if the judge formed the intention to alter circumstances before his birth?" – is completely different than the question of what would happen if the

judge formed the intention to raise his hand. Similarly with violating the laws of physics.

In fact, van Inwagen's entailment principle is false in general. Here is a mundane counterexample. Imagine a linear algebra student who did not study enough for her test. She is asked to compute the inverse of a 6×6 matrix. She cannot compute the inverse. (Even if she formed the intention to invert the matrix, she still would not successfully do it.) But for each individual step of the Gauss-Jordan elimination algorithm, she can perform that individual step. (If she formed the intention to e.g. replace row 2 of the matrix with the sum of rows 1 and 2, she would easily do it.) She can perform each step of the algorithm, and performing all of the steps entails inverting the matrix, but she can't invert the matrix.

For an even simpler counterexample, suppose the password for access to the United States' nuclear weapons is "09384294". A Soviet spy wishes to gain access, but does not know the password. The spy cannot enter the correct password. (Even if she thinks, "I shall enter the correct password," she still fails to enter the correct password.) But she can enter "09384294", and doing so entails entering the correct password.

Thus, van Inwagen's argument for the incompatibility of free will and determinism is invalid. The point that van Inwagen seems to miss is that even if X and Y logically entail each other, determining the truth of "S can do X" requires consideration of one counterfactual world (the world where S thinks to herself "I shall do X,") whereas determining the truth of "S can do Y" requires consideration of a different counterfactual world (the world where S thinks to herself "I shall do Y.")

I'll end by responding to van Inwagen's discussion of the conditional analysis. He correctly observes the potential threat posed by the conditional analysis to his argument: When his premises are reconsidered with the conditional analysis in mind, it might become clear that some premise is false. He considers one such premise (not the premise to which I objected) and argues that the conditional analysis leaves it unharmed, because either the premise is still true (it certainly still seems true), or else his argument in favor of the premise becomes an argument against the conditional analysis. He then says, "The same dilemma confronts the conditionalist if he attempts to show, on the basis of the conditional analysis, that any of the other premisses of the argument is false." [2, page 227] I suppose that I do face that dilemma, but I believe I have overcome it. Indeed, his defense of his entailment principle is an argument against the conditional analysis. But it is an invalid argument against the conditional analysis, as I believe I have convincingly argued already.

Have I saved compatibilism from van Inwagen's attack? It depends on what you mean. I do not pretend to have argued that free will and determinism are compatible; maybe van Inwagen's argument can be patched up. But, at least, I do believe that I have uncovered an unreasonable assumption of van Inwagen's about how the word "can" works. He wanted his assumptions to be "basic and evident enough to be data that an

analysis of this concept must take account of," so that "any analysis on which these claims did not 'come out true' would be for that very reason defective." [2, page 216] On the conditional analysis, his entailment principle does not come out true. The conditional analysis may be defective, but not for that very reason.

References

[1] Van Inwagen, Peter. "The incompatibility of free will and determinism." *Philosophical studies* 27.3 (1975): 185-199.

[2] Pereboom, Derk. *Free Will.* Indianapolis, IN: Hackett Pub., 1997. Print.