A Refutation of the Lewis-Stalnaker Analysis of Counterfactuals

Marcus Arvan

University of Tampa

marvan@ut.edu

Abstract: The standard philosophical analysis of counterfactual conditionals—the Lewis-Stalnaker analysis—analyzes the truth-conditions of counterfactuals in terms of nearby possible worlds. This paper demonstrates that this analysis is false. §1 shows that it is a serious epistemic and metaphysical possibility that our "world" is a massive computer simulation, and that if the Lewis-Stalnaker analysis of counterfactuals is correct, then it should extend seamlessly to the case that our world is a computer simulation, in the form of a *possible-simulation semantics*. §2 then shows, however, that a Lewis-Stalnaker-style possible-simulation semantics clearly fails as an analysis of the truth-conditions of counterfactuals in two types of simulated worlds: Humean Simulations and Necessitarian simulations. §3 then considers and answers several objections. Finally, §4 draws several skeptical, but compelling lessons about counterfactuals from the argument.

Key words: counterfactuals, conditionals, semantics, simulations, meaning.

The standard philosophical analysis of counterfactual conditionals—the *Lewis-Stalnaker analysis*—analyzes the truth-conditions of counterfactuals in terms of nearby possible worlds.[1] Although Lewis, Stalnaker, and others[2] have disagreed significantly over details of the analysis—in particular, how to understand closeness of possible worlds, as well as the nature of possible worlds themselves—and although some have challenged the approach of understanding counterfactuals in terms of possible worlds[3], the Lewis-Stalnaker analysis is widely accepted and utilized in philosophy today. This paper demonstrates this analysis to be false.

§1 affirms, on the basis of recent work by physicists and philosophers, that it is a serious epistemic and metaphysical possibility that our "world" is a massive computer

[1] See Stalnaker (1968) and Lewis (1973, 1979).
[2] See e.g. Bennett (2003): §§10-19.
[3] See e.g. Jacobs (2010)

simulation, and that even our world is not actually a simulation, it could have been one (i.e. our world could be simulated).[4] §1 then argues that if the Lewis-Stalnaker analysis of counterfactuals is correct, then it should extend seamlessly to the case that our world is a computer simulation, in the form of a *possible-simulation semantics*. §2 then shows, however, that a Lewis-Stalnaker-style possible-simulation semantics clearly fails as an analysis of the truth-conditions of counterfactuals in simulated worlds. §2.1. argues, first, that that there are two possible types of simulations: (A) "Humean simulations" in which objects, properties, and events are pre-programmed into the simulation completely independently of one another, without falling under any general, computationally-encoded "simulation laws"; and (B) "Necessitarian simulations" in which objects, properties, and events *are* governed by general laws written into the simulation's program. §2.2. then shows that a Lewis-Stalnaker possible-simulation semantics is clearly false for Humean simulated worlds, and §2.3. shows that it is also clearly false for Necessitarian simulations. I conclude, as such, that the Lewis-Stalnaker semantics for counterfactuals is false. Because the analysis must provide suitable truth-conditions for counterfactuals in simulations, but it plainly fails to do so, it cannot be a correct theory of the semantics of counterfactuals. §3 then considers and answers several objections to the argument. Finally, §4 draws several skeptical lessons from the argument: namely, namely, that setting aside some counterfactuals we can know the truth-value of merely on the basis of actual sequences of events, (1) we must know what kind of world (Humean or Necessitarian) ours is in order to know which counterfactuals are true in our world, (2) we must know which kind of world ours is to know what the correct semantics for counterfactuals is (since, as I argue, Humean and Necessitarian realities entail different semantics), and finally (3) since we cannot possibly know which type of world ours is (Humean or Necessitarian) from within it—we could only know what type of world ours is if we were "freed" from it and perceive its functional architecture from the outside (much as "Neo" is freed from "The Matrix" and able to view its code in *The Matrix* films)—we cannot know precisely which counterfactuals or semantics for counterfactuals are true of our world. Although I recognize that these implications may seem "counterintuitive" to some, I submit that the world is under no

[4] I place 'world' in scare-quotes here to allude to the fact that if our world is a simulation, it is a *simulated* world—one that exists as a computer program within some kind of broader concrete (meta-)reality.

obligation to be intuitive, and moreover, that once we understand that our world may be a simulation, the notion that we cannot know which counterfactuals are true of it is not counterintuitive at all.

§1. Our World Might Be a Simulation, and Why It Matters for Counterfactuals

The idea that we may be living in a computer simulation is gaining momentum both in philosophy and in physics. A little over a decade ago, Nick Bostrom argued that it is probable that we are probably living in a computer simulation.[5] A couple of years later, David Chalmers argued that the hypothesis that we are living in a computer simulation should be understood not as an epistemic hypothesis calling into question knowledge of the external world, but rather a metaphysical hypothesis about the nature of the external world.[6] More recently, a number of physicists have proposed a numerical-functional analysis according to which our universe might be a simulation.[7] Finally, Marcus Arvan has recently argued that a new version of the simulation hypothesis, the Peer-to-Peer (P2P) Simulation Hypothesis, is not only entailed by several serious hypotheses in philosophy and physics, but also promises to provide a unified explanation of many observed features of our world that currently lack such an explanation: quantum indeterminacy, superposition, wave-particle duality, and entanglement in physics[8], as well as the mind-body problem, problem of free will, problem of personal identity, and problem of time's passage in philosophy.[9]

For all we know, then, our "world" is metaphysically identical to a simulation. Secondly—and this is important—even if our world turned out not to be a simulation, the arguments that Bostrom, Chalmers, and others have given indicate that our world *could* have been a simulation. Both points are directly relevant to the semantics of counterfactuals. Counterfactuals clearly hold for simulated worlds just as much as they hold for non-simulated worlds. For instance, when I play a game of Halo or Call of Duty—two common online simulations—there are all kinds of true counterfactuals for the game. Here is one: if I *were* to shoot an enemy player X number of times, that player in the game

[5] Bostrom (2003).
[6] Chalmers (2005).
[7] See e.g. Beane et al. (2012).
[8] See Arvan (2014): 437-445. Also see Arvan (2013): §§III-IV.A-B.
[9] See Arvan (2012): §IV.C-F.

would "die." Here is another: if I were to walk my playable character into a teleporter in Halo, my character would appear on the other side. Accordingly, if our world is a simulation, surely there are true counterfactuals for *it* as well.

Notice, next, what the standard Lewis-Stalnaker analysis of counterfactuals entails for simulations. If the our world is (or were) a simulation, then on the Lewis-Stalnaker account the semantics for counterfactual conditionals for our world must be given by other nearby simulations, in the form of a *possible-simulation semantics*. We can see that this is a straightforward implication of the Lewis-Stalnaker theory as follows. First, if our world is a simulation, then (as we saw above) there are true counterfactuals for that simulation. If, however, there are true counterfactuals for the simulation we live in, then—on the Lewis-Stalnaker analysis—the truth-conditions for those counterfactuals is given by the closest nearby possible worlds. If, however, our world is a simulation, then the closest nearby worlds—on any standard metric of world-closeness (e.g. Lewis' metric, etc.)—are *also* worlds in which our world is simulation (viz. possible worlds in which we are not living in a simulation would be very dissimilar to the actual world). Thus, *if* the Lewis-Stalnaker analysis of counterfactuals is correct, then the semantics for counterfactuals in simulated worlds should be given by *nearby simulated worlds*. As such, by contraposition, if the Lewis-Stalnaker analysis fails for simulated worlds—if it provides incorrect semantics for counterfactuals in simulated worlds—then the Lewis-Stalnaker analysis of counterfactuals is incorrect.

§2. The Refutation

We have just seen that, for all we know, our world is metaphysically identical to a computer simulation. What, though, is a "simulated reality" exactly? We need not provide necessary or sufficient conditions in order to proceed fruitfully. For our purposes, we can simply define the notion of a simulation intuitively—as *the kinds of computer programs we have already created encoding virtual environments*, i.e. functional analogues to the kinds of things that exist in our world (e.g. functional analogues of rocks, trees, cars, people, etc.).

§2.1. Two Types of Simulations

Philosophers have long debated whether our world is broadly "Humean" or "Necessitarian" in nature. The Humean view of reality holds that there is "nothing to reality except the spatio-temporal distribution of local natural properties"[10]—that is, it holds that reality is a *mosaic* of *logically and metaphysically independent* objects and properties that, as such, bear *no necessary connections* to one another and could, in principle, be *arranged or rearranged* in any logically possible spatio-temporal distribution. There is a simpler way to put this. According to the Humean view, reality is *literally* a mosaic in roughly the same way that a painting is. No blob of paint on a painting bears any necessary relationships to any other. Instead, a painting *just is* blobs of paint in different spatial positions on a canvas—blobs which may be in some configuration (e.g. depicting the Mona Lisa), but which could exist in and endless variety of other configurations. Another helpful way to understand the Humean view of reality is on the analogy of a film strip. Nothing on any particular frame of a filmstrip, or the order of the frames, *has* to be the way it is. Consider, for instance, a film of a person swinging a tennis racket. Each frame of this filmstrip would contain an image of a person in a slightly different position, such that when the filmstrip is played, we see a *moving image* of the person swinging the racket. Although the images show this, however, *they didn't have to*. One could erase the film and encode different images entirely: a film of someone chewing gum, perhaps. Alternatively, one could cut up the film and paste its frames back together in a completely different order. The Humean view of reality says that reality is analogous to this: that there are no *necessary connections* between, say, protons and electrons, and that the world could have been rearranged so that the same objects—the same protons and electrons—behaved in very different ways.

In sharp contrast, the Necessitarian view holds that reality is *not* a mosaic of logically and metaphysically independent parts that could that can be arranged in any manner whatsoever.[11] Instead, on the Necessitarian view, objects and properties in the world have *powers* that are essential to them, and are *necessarily connected* to other objects and properties. According to the Necessitarian view, electrons *necessarily* repel positively charged particles. As such, in contrast to the Humean view, which holds that objects and properties could exist in *any* spatio-temporal configuration, the Necessitarian view holds that certain types of configurations of objects and properties are metaphysically

[10] Weatherson (2014): introduction.
[11] Again, see Jacobs (2010).

impossible: electrons, for instance, due to their essential nature, could not *possibly* exist in certain spatio-temporal positions relative to positively charged particles.

As we will now see, both metaphysical conceptions of the world can be *simulated*, and this can teach us important lessons about the semantics of counterfactuals.

§2.1.1. Humean Simulations

As we have just seen, according to the Humean picture of reality, our world is something like a filmstrip. In a filmstrip, there is not only (1) no essential connection between objects or properties within any single frame of film (each "pixel" bears no necessary relations to any other "pixel"; one can paint in each pixel however one likes); (2) there are also no essential connections between different film frames (one could, in principle, chop up and rearrange all of the individual frames of the filmstrip in whatever order one likes). According to the Humean conception of reality, none of reality's parts bear any *essential* or *necessary* relationships to any other. Each space-time point in reality is *metaphysically distinct* from any (and every) other.

Interestingly, it is possible to program simulated worlds that are Humean in this sense. Here is how. A meticulous programmer would simply have to program *every pixel* in every "frame" of the simulation independently, one by one. So, for instance, consider my visual field right now. I see a laptop computer, a table, a person in the background, cars into the parking lot. A meticulous computer programmer could have programmed each of those pixels into the program—each and every *point* of every object—independently of every other, as a kind of "atomic fact" in the simulation (viz. the programmer would program code of roughly the following sort "pixel#1/time *t*=solid, brown, etc.", pixel#2/time *t*=liquid, blue, etc.", etc.). Such a simulation would, in other words, not have any general governing "laws" programmed into it (e.g. no virtual law of gravity, etc.). Rather, the "law-like" behavior of all objects in the simulation (e.g. dropped objects accelerating toward the ground) would simply be the result of the programmer manually programming each pixel in every subsequent "frame" of the simulation, such that when the simulation is played—producing a "moving picture"—objects behave in regular ways. Creating this kind of simulation, pixel by pixel, would of course be an extraordinarily arduous process—and indeed, animated movies and computer-animated movies were for many years created in just this kind of manner (with each pixel in every "frame" of the film painted or programmed in manually).

Given that, for all we know our world is a simulation, our world may be a *Humean* simulation of just this sort. Our world, that is, may not have any law of gravity explicitly programmed into it (no code of the form, "If x is an object, x will fall at $9.8m/s^2$, etc). Instead, if our world is a Humean simulation, the *fact* that we observe objects fall at $9.8m/s^2$ is generated instead merely by *completely independently-coded* "simulation-frames" arranged in such a manner that when the simulation is executed, the series of moving frames generate objects regularly accelerating to the ground at $9.8m/s^2$**.** In other words, although laws of nature—such as the law of gravity—would not actually be *explicitly encoded* into the program of our world, it would *look* to us as though it is.

§2.1.2. Necessitarian Simulations

Now consider a second type of simulation, one in which computational laws—e.g. simulated laws of gravity, electromagnetism, etc.—*are* explicitly encoded into the simulation's program (viz. computer code of roughly the following form: "if x is an object in the simulation, x will accelerate downward at $9.8m/s^2$). Most videogame nowadays have this sort of functional structure: their simulated world "game engines" have simulated physical laws (laws of gravity, speed, etc.) explicitly programmed into them. Call this a Necessitarian simulation.

Notice that a Necessitarian simulation, as such, is very different than the kind of Humean simulation discussed earlier. Whereas every pixel—every object and property—in a Humean simulation bears *no necessary connection* to any other (in that each pixel is programmed in individually, and independently of every other), in the kind of Necessitarian simulation we are now describing this is not the case: there are *computational laws* encoded in the latter which govern how different objects and properties in the simulation relate to one another. If, for instance, there is an object represented at point A at time *t* in the simulation and the programmer codes in the law, "All objects at any point X at any time *t* in the simulation will accelerate in manner Y at time *t+1*", then this law will *govern* what happens to the object at point A. In other words, unlike a Humean simulation, where *no necessary connections between objects or events* are encoded into the simulation, in a Necessitarian simulation there *are* necessary connections between objects and events explicitly encoded into the simulation's program to *govern* how objects and properties related to one another.

There is an important point to notice here that we will return to later: namely, that Humean and Necessitarian simulations would be *indistinguishable from the inside* (i.e. from the frame of reference of any observer within them). The reason for this is simple. Both types of simulations could, in principle, encode *precisely the same series of events* (e.g. my sitting here typing the words I am typing this very moment). A Humean simulation would simulate the law of gravity in virtue of the programmer plotting out every individual object, in each successive frame of the simulation, such that when the Humean simulation is played, objects appear to accelerate at a rate of 9.8m/s^2. A Necessitarian simulation would produce the very same observations to observers in the simulation, but through a very different means: namely, computational governing laws explicitly coded into the simulation that determine how objects in the simulation move. These genuine, functional differences between a Humean and Necessitarian simulation would, as such, be inaccessible and unknowable to any individual within such a simulation, and could only be known or accessed from a standpoint outside of the simulation—for instance, from the point-of-view of a programmer who can *see* whether the simulation is programmed in a Humean or Necessitarian manner.

§2.2. Why Lewis-Stalnaker Semantics is False for Humean Simulations

As we saw earlier, if the Lewis-Stalnaker analysis of the semantics of counterfactuals is correct, the truth-conditions for counterfactuals in simulations should be given by what occurs in nearby possible simulations. However, does the Lewis-Stalnaker analysis provide a correct semantic analysis for counterfactual conditionals in "Humean simulations"? The answer, as we will now see, is clearly no.

In order to determine whether the Lewis-Stalnaker analysis provides a correct analysis of truth-conditions of counterfactuals in a Humean simulation, we should first reflect carefully on precisely what the Lewis-Stalnaker analysis is supposed to do. Proponents of the analysis—namely, Lewis, Stalnaker, and other semanticists—have always understood the analysis as providing the truth-conditions for counterfactuals: that is, a theory of the conditions that *actual* makes counterfactual conditionals true or false.

This is important because, as we will now see, people living in a Humean simulation might *think* certain counterfactuals are true when in fact they are not. Allow me to explain.

Consider a counterfactual about our world that we would ordinarily consider to be true. Here Jones is standing before an electrified wire. Jones has not yet touched it. Yet, since it is electrified, we are apt to say that *if Jones had touched it, Jones would be shocked*. It is natural enough to think this counterfactual is true, obviously, given the kinds of regularities we have experienced in our world to date (i.e. people getting shocked when they touch live electrical wires). However, as we will now see, whether the counterfactual *is in fact true* cannot be ascertained by anyone in a Humean simulation. Allow me to explain.

Suppose you were a programmer of Humean simulations, and you programmed a variety of different such simulations. Suppose the first Humean simulation you programmed, S1, is a simulation in which:

1. Anytime anyone touches a live wire (X occurs), they are in fact shocked (Y occurs).
2. Anytime someone does not touch a live wire (X does not occur), they are not shocked (Y does not occur).
3. Jones almost touches a live electrical wire at *t*, but does not and does not receive a shock.

Now suppose that the second Humean simulation you program, S2, is just like S1 in every way accept that instead of (3), (4) holds:

4. A counterpart of S1's Jones, Jones*, touches the wire at *t* and receives a shock.

Finally, suppose that in a third Humean simulation, S3, you program neither (3) nor (4) but instead (5):

5. Another counterpart of S1's Jones, Jones**, touches a live wire at *t* but does not receive a shock.

Now consider the counterfactual conditional, "If Jones had touched the wire at *t*, he would have been shocked." The Lewis-Stalnaker analysis holds that the counterfactual is *true* at S1 because S2 is the most similar counterfactual worlds in which (a counterpart of) Jones touches the wire (e.g. S2), Jones is shocked. However, is this really the right way to understand that counterfactual's truth-conditions?

Consider how a programmer would think about counterfactuals for these different simulations. Would a programmer be at all tempted to say that the counterfactual, "If Jones had touched the wire, he would have been shocked" is true in *S1* because a *counterpart* of Jones who touches the wire in the most "similar" simulation(s) (e.g. S2) is shocked? No programmer would ever say or assent to such a thing, and for an obvious reason: such a programmer would recognize that nothing that happens in one Humean simulation *depends in any way* on what happens in any other. Indeed, in conversation, programmers often explicitly *index* counterfactual claims to the specific programs they have written. If you were to ask a programmer of simulation S1, "what would happen if Jones touched the wire at *t* in S1?", they would say *not* refer to "nearby simulations" in giving an answer. Rather, they would say, "There's no answer to that question. I programmed S1 so that Jones *doesn't* touch the wire, so there's no fact of the matter of what would happen *in S1* if he touched the wire there." "Now," they might add, "if I had programmed S1 differently—if I had programmed it as I did S2—*then* it would be true in S1 that if Jones touched the wire, he would have been shocked. Similarly, if I programmed S1 like S3, then it would be true in S1 that if Jones touched the wire, he wouldn't be shocked." Finally, the programmer might add, "In other words, before I can tell you what *would* happen in S1 if such-and-such were the case, I need to specify how S1 is programmed. If S1 were programmed thusly, this would happen. If it were programmed like S2, then *that* would happen. Etc."

Now, it might be objected to this that, in making this argument, I have appealed to the programmer's perspective, not the perspective of inhabitants living within any of the simulations described. This, however, is by design. For the point of adopting a programmer's point-of-view is to draw attention to certain facts that the programmer knows which are directly relevant to the actual truth-conditions of counterfactuals: namely, the facts that (1) what happens in one Humean simulation has no bearing whatsoever on the goings-ons of any other simulation (no Humean simulation depends in any way on any other), and (2) each simulation's own program dictates would happen in that simulation if it were "played." The point of adopting the programmer's perspective is to draw attention to the fact that while the inhabitants of a Humean simulation might think that certain counterfactuals are true of their world (viz. "If Jones touched the wire he would be shocked"), the reality is that whether that counterfactual is actually true depends entirely on features of their world's programming to which they do not have any access (i.e. if their world is a Humean simulation, then what would happen in their simulation is entirely given by what it was in fact programmed to do).

§2.3. Why Lewis-Stalnaker Semantics is False for Necessitarian Simulations

Now turn to Necessitarian simulations. A Necessitarian simulation is, if you recall, a simulated reality whose computational architecture (e.g. computer code) *does* involve explicitly-encoded "laws of nature" (e.g. lines of code to effect of, "For all objects, *x*, if *x* is in conditions *y* at time *t*, plot *x* in conditions *z* at *t+1*). Again, such simulations are common. Programmers of online simulations such as Halo, Call of Duty, etc., explicitly code "game physics" into these simulations (simulated rules of gravity, simulated rules to govern lighting/reflectivity, etc.). However, does the Lewis-Stalnaker analysis provide a correct semantic analysis for counterfactuals in such simulations? Here, just as with Humean simulations, the answer is clearly no.

Consider once again the perspective of a programmer. Suppose you had programmed several Necessitarian simulations. In simulation S1, you program a law of nature to the effect of, "For all objects x, x will be shocked if and only if x comes into contact with any electrified object, y [where electrified objects are defined according to some other lines of computer code]." Suppose, next, that you programmed S1 such that it is inevitable that Jones does *not* touch an electrified wire at time *t*. It is clearly true, at S1, that if Jones *had* touched the wire, he would have been shocked—but notice: there are no reasons to analyze the truth of this counterfactual in terms of anything other than the law explicitly programmed into S1. The counterfactual, "If Jones had touched the wire, he would have been shocked", is true at S1 *not* because of anything that happens (or might happen) in any "nearby simulation." It is true simply because Jones is an inhabitant of S1 *and S1's computer code explicitly encodes a functional rule to the effect of: if he were to touch the wire, he would be shocked* (for all objects, x, x will be shocked if and only if x comes into contact with an electrified object).

So, Lewis-Stalnaker semantics is false for Necessitarian simulations as well. The truth-conditions for counterfactuals in any given Necessitarian simulation are given simply by the computational "laws" and initial conditions programmed into that very simulation. The laws and initial conditions of any given Necessitarian simulation determine what *would* happen in that simulation. Now, of course, different Necessitarian simulations might be programmed with very different laws, such that we might say of one simulation, S1, that had it been programmed differently (viz. S2), different things would happen. But in that case, too, it is not the Lewis-Stalnaker analysis that gives a correct semantic analysis. For notice: if we were to say that different things would happen in S1 if it had S2's programming and initial conditions (which is true), we are still not evaluating counterfactuals in S1 in terms of "nearby possible worlds"; rather, we are *stipulatively* picking out a relevant counterpossible simulation (S2) and deriving what would happen in S1 if it (S1) had that counterpossible simulation's *precise* laws and initial conditions.

There is another way to put this: namely, that just as with Humean simulations, a programmer of a Necessitarian simulation would *index* the truth of counterfactual conditionals to each simulation's programming—in this case (the Necessitarian case), however, to each simulation's *governing laws* and initial conditions. A programmer who programmed a Necessitarian simulation in which Jones touches a wire and is shocked (following that simulation's laws and initial conditions) would say, of that simulation, "If Jones touched the wire, he would be shocked." However, if you were to ask the same programmer the alternative question, "What would happen if Jones hadn't touched the wire?", the programmer would answer, "*That depends on which alternative programming we're talking about*. If I had programmed the laws and initial conditions to lead to Jones' not touching the wire to lead to his not being shocked, then he would not get shocked. However, if I had programmed the laws and initial conditions to lead to Jones not touching the wire *but experiencing a shock*, then the opposite would be true."

§2.4. Conclusion: The Lewis-Stalnaker Analysis is False

We have seen that if the Lewis-Stalnaker analysis of counterfactuals were true, it would have to extend to simulated worlds in the form of a possible-simulation semantics invoking "nearby possible simulations." We then saw, however, that there are two very different types of simulated realities: Humean simulations, which functionally realize objects, properties, and states of affairs (simulated tables, chairs, electrons, etc.) without any governing "laws", and Necessitarian simulations, which realize objects properties, and states of affairs utilizing governing laws (e.g. simulated laws of gravity, etc.). Finally, we saw that the Lewis-Stalnaker analysis fails to provide the correct truth-conditions for counterfactual conditional in both types of simulations. It fails for Humean simulations for the simple reason that what happens in one Humean simulation depends in no way on what happens in any other Humean simulations; and it fails for Necessitarian simulations for the simple reason that a Necessitarian simulation's governing laws and initial conditions—rather than "nearby worlds"—specify which counterfactuals are true for it. Thus, I submit, the Lewis-Stalnaker analysis is false.

§3. Objections, and Replies

I suspect several objections to the purported refutation just provided. Allow me to raise and respond to what I take the most likely objections to be.

Objection #1-Nothing More Than a Standard Skeptical Argument[12]: Some readers might object that it is not exactly a surprise—nor a problem—that the Lewis-Stalnaker analysis fails for simulations. After all, one might think, if we are in a simulation, just about *everything* we think we know—about tables, chairs, cars, people, etc.—is false: there are no tables, chairs, cars, or people, only simulated ones. In other words, some readers say, I have not actually motivated an argument against the Lewis-Stalnaker possible-worlds *semantics* for counterfactuals. I have simply motivated a general *skeptical argument* calling into question whether we *know* the Lewis-Stalnaker analysis applies to the 'world' in which we live (in much the same way that simulation/skeptical arguments call into question our knowledge of *everything* outside of our own minds).

Reply: This objection misunderstands the argument. Simulated trees, rocks, and people may or may not be "real" trees, rocks, or people (though people like Chalmers and Arvan have suggested that we should identify them as real in every relevant sense—since, if we are in a simulation, they are the (simulated) things we interact with one a day-to-day basis). As such, we may not *know* whether "real" trees, rocks, or people exist. But, for all that, if we are in a simulation, *simulated trees, rocks, and people* exist—and there are (as we have seen) *true counterfactuals for these simulated things*. But now a correct semantics for counterfactuals purports to give the truth-conditions not just for some counterfactuals on some metaphysical or epistemic assumptions: a correct semantics for counterfactuals *should give the correct truth-conditions for counterfactuals simpliciter*. And this point of this paper is that there are genuine, true counterfactuals (counterfactuals about simulated

[12] I thank an anonymous reviewer for raising this concern.

objects) that the Lewis-Stalnaker analysis gives the wrong truth-conditions for. It cannot, therefore, be *the correct semantic theory* of counterfactuals, as there are some circumstances where it gets their meaning and truth-conditions incorrect.

Objection #2-Why the Detour into Simulations?[13]: Some readers might wonder whether my argument's focus on simulated worlds is beside the point, as the challenge I am raising for the Lewis-Stalnaker analysis is based on the *nature* of Humean and Necessitarian realities *per se*, simulations or not. After all, isn't my argument that the Lewis-Stalnaker analysis is false for Humean worlds based on the point that *no* Humean world—simulated or otherwise—depends in *any* way on any other Humean world? And isn't my second argument then that the Lewis-Stalnaker analysis fails for Necessitarian worlds because it is the *governing laws* of those worlds that make counterfactuals true, not Lewis-Stalnaker world-similarity? What, then, is the point of talking in terms of simulations?

Reply: In one sense, this is correct—and it speaks again to the first objection (the objection that my argument is merely an epistemic one calling into question our *knowledge of whether* the Lewis-Stalnaker analysis is correct). My argument, indeed, does not rely essentially on the simulation hypothesis. It is based on the very nature of Humean and Necessitarian *worlds* (simulated or otherwise). Nevertheless, the discussion of simulated worlds is helpful for two related reasons

First, good philosophical arguments are intuitively forceful—and the simulation hypothesis renders the problems the nature of Humean and Necessitarian worlds raise for the Lewis-Stalnaker analysis *vivid and intuitively forceful*. Proponents of the Lewis-Stalnaker analysis have long recognized the worry that there might be something wrong with analyzing modal notions for one world in terms of what happens in other worlds. Indeed, the most well-known objection to possible-worlds modal semantics—the so-called "Humphrey Objection"—is very similar (but not identical!) to mine. The Humphrey Objection is this: how can what a *counterpart* of me does in another possible world be at all relevant to what *I* might have done in this world?[14] As Sider writes, "Kripke's complaint in Naming and Necessity was that while Hubert Humphrey cares very much that *he* might

[13] I thank another anonymous reviewer for raising this concern.

[14] See Kripke (1972): 45.

have won the 1968 U.S. presidential election, [Humphrey] "could not care less whether someone else, no matter how much resembling him, would have been victorious in another possible world."[15]

The objection this paper defends is similar but not identical to the Humphrey Objection. The Humphrey Objection is motivated by the worry that an individual in one possible world (Humphrey) *doesn't care* about what happens to counterparts of him in another possible world. People like Lewis have responded effectively enough to this worry, I think, by claiming that *if Humphrey understood counterpart theory properly*—namely, if Humphrey understood that counterparts of him in other worlds *represent (by proxy) modal properties that are true of him*—then he would understand that he *should* care about his modal counterparts (as they represent *his* modal properties). My point, using simulations, is not to simply raise the Humphrey objection again, but rather to make especially vivid that there are *general facts* about Humean and Necessitarian worlds that make the Lewis-Stalnaker analysis of counterfactuals false for them. Now, it is certainly true that, in principle, I needn't have utilized simulated worlds to make the point. Still, invoking simulated worlds has a point: it makes the argument *concrete, in a way that engages with our ordinary-everyday experience of the world*. Talk about "possible worlds", after all, is incredibly abstract, artificial, and unique to philosophy. It is only *philosophers* who talk explicitly about possible worlds—and because possible worlds are just that (*possible* worlds), we do not come into contact with them in any intuitive way. Indeed, many philosophers have argued, *contra Lewis*[16], that it is wrongheaded to understand possible worlds as concrete objects. Lewis, of course, was a modal realist: he believed that possible worlds are *concrete alternative universes*. Most philosophers, on the contrary, reject modal realism, and for obvious enough reasons: we appear to have no concrete, observational evidence for the existence of alternative universes. Consequently, many philosophers have argued that we should understand possible worlds as abstract objects, fictional objects, etc. But, of course, once we start talking about abstract objects, fictional objects, etc., it is hard to have firm, well-grounded philosophical intuitions (for instance, it seems intuitive enough, I suppose, that if possible worlds are fictions, Hubert Humphrey *should* care about

[15] Sider (2006): 1.
[16] Lewis (1986).

other possible worlds—for they are fictions that *could have been true of him*). *Simulated* worlds, on the other hand, are not abstract or fictional. They are concrete objects, and it is my contention that by *looking* at them—by looking at their concrete features—we can see, in down-to-earth terms, what is wrong with the Lewis-Stalnaker analysis.

Second, simulated worlds not only make the argument vivid and concrete, giving us concrete reasons for believing the Lewis-Stalnaker analysis to be false—again, each type of simulated world has *functional-dispositional properties* (Humean worlds have no governing laws, Necessitarian worlds do) that make the Lewis-Stalnaker analysis ill-suited to analyze counterfactuals correctly. Simulated worlds also enable us to sidestep questions about the metaphysical nature of "possible worlds." We need not figure out whether possible worlds are concrete objects (as Lewis argues), whether they are abstract objects, fictions, or whatever—things which we may or may not have clear intuitions about. Because simulated worlds are concrete objects—we can create, and have created, many simulated realities (at least on a small scale)—and (as we saw earlier) the Lewis-Stalnaker analysis plus any standard account of world similarity metric would have us analyze what is counterfactually true in one simulation in terms of what is true in "nearby simulations", it follows, on the Lewis-Stalnaker, that whatever "possible worlds" ultimately turn out to be (abstracta, concreta, fictions, etc.), actual simulations—the kinds we have created—are material instantiations of those things: if possible worlds are fictional entities, actual simulations are material realizations of some such fictions; if possible worlds are abstract objects, actual simulations are material realizations of the possibilities those abstracta represent; etc. In other words, simulations enable us to do the following: whatever possible worlds are (fictions, abstracta, etc.), we can investigate what is true of possible worlds by investigating simulations (since simulations embody possibilities that abstract/fictional/etc. possible worlds represent). Simulated worlds enable us, in other words, to evaluate the Lewis-Stalnaker analysis from a perspective that is "standpoint neutral" with respect to the ultimate nature of possible worlds (something which I think is clearly desirable: a good argument should beg as few questions as possible). In conclusion, then, the argument's reliance on simulated worlds is not incidental to it. While it might be possible to make something like my argument without appeal to simulated realities, the appeal to simulated realities (A) makes the argument helpfully vivid and concrete, in a manner that also enables us to (B) set aside distracting, and potentially confounding, questions about the nature of possible worlds.

Objection #3-A Bare Appeal to Intuitions?: A second objection I anticipate to my argument is that I have appealed (repeatedly) to "intuitions that a programmer would have" about the truth-conditions of counterfactuals in different types of simulations. First, I argued that a programmer of Humean simulations would "not be tempted at all" to evaluate counterfactual conditionals for one simulation in terms of any other. Similarly, I argued that a programmer of Necessitarian simulations would analyze the semantics of counterfactuals for those "worlds" merely in terms of each simulation's respective laws and initial conditions. These, however, are *mere intuitions*. How can I say "what a programmer would say"? Moreover, why should it *matter* "what a programmer would say"? What we want to know is what the truth-conditions of counterfactuals are—and given that the Lewis-Stalnaker theory is otherwise plausible and firmly entrenched in philosophical discourse, shouldn't we insist that a "refutation" of the theory be based on something better than "programmer intuitions"?

Reply: The argument against the Lewis-Stalnaker analysis is not merely based on intuitions. The point of appealing to the "programmer's point-of-view" is to bring out objective features of different types of simulations that reveal problems with the Lewis-Stalnaker analysis. The Lewis-Stalnaker analysis says that we should understand the truth-conditions of counterfactuals for any given world in terms of what happens in nearby possible worlds. Yet, we saw, first of all, that *no Humean simulation depends in any way on what happens in any other*. Because there is no dependence between what occurs in one Humean simulation and what occurs in another, there are no philosophical grounds for analyzing what would counterfactually happen in one in terms of what would happen in the other. In other words, once we reflect on what Humean simulations are—their having no functional or causal dependence on each other whatsoever—we see that a programmer (and indeed, the rest of us) *have good reasons* to reject the Lewis-Stalnaker analysis for Humean simulations. Given that what happens in one Humean simulation has no bearing on what happens (or would happen) in any other, the Lewis-Stalnaker analysis has to be false for Humean simulations. Similarly, the point in the discussion of Necessitarian simulations was not to invoke bare, unsupported intuitions about counterfactuals in those simulations. Rather, the point was to draw attention to objective facts about Necessitarian simulations that make the Lewis-Stalnaker analysis plainly inadequate for counterfactuals within them. Since any given Necessitarian simulation has *governing laws* encoded into its functional architecture, what would happen in any given Necessitarian simulation is—objectively—not a matter of what happens in any "nearby simulation" but rather something that can be *derived directly from a statement of the simulation's laws and initial conditions*.

Objection #4-absurd implications?: My argument has been that in order to know which counterfactuals are true of a given simulation, we have to know—objectively—what kind of simulation (Humean or Necessitarian) it is. So, for instance, I argued that *if* our reality is a Humean simulation, then the truth-value of any given counterfactual—for example, the truth-value of the counterfactual, "If Jones hadn't touched the wire, he wouldn't have been shocked"—depends entirely on what actually happens in our world. I argued, in particular, that if Jones touches the wire in our world (and our world is a Humean simulation), then the counterfactual, "If Jones hadn't touched the wire, he wouldn't have been shocked", *has no determinate truth-value* (since, in our simulation, the antecedent to this counterfactual—Jones' not touching the wire—cannot be satisfied). Similarly, I argued that in order to know the truth-value of counterfactuals in Necessitarian simulations, one has to know that it is a Necessitarian simulation (since, if I am correct, the semantics for counterfactuals in Necessitarian simulations are given by their governing laws and initial conditions). But, the objection goes, this is clearly wrong-headed. The notion that we must know what type of world we live in before we can know which counterfactuals are true is, quite simply, bizarre. We know which counterfactuals are true in our world, simulation or no. It's true, for instance, that if I threw this computer to the ground right now and hit it with a sledgehammer, then I would lose all of my hard work on this article. It's true, if anything is, that if I were to touch a live electrical wire, then I would be shocked. And the Lewis-Stalnaker analysis gives the correct truth-conditions for counterfactuals such as these.

Reply: Although we of course think we know which counterfactuals are true of our world, if our world is a simulation (and it may be), or alternatively, if it were a simulation, there is not a programmer in the world—nay, not a single reasonable person in the world!—who would say we could know the truth-value of counterfactuals merely on the basis of our experiences. Allow me to explain why. Although of course it seems obvious that if I were to hit this computer with a sledgehammer, I would smash it to bits, any programmer worth their salt would tell you that, *if our world is a simulation, you cannot know this is true—at least not unless and until you know what kind of simulation we live in*. Here is why. Suppose you were to ask a programmer, of any simulation whatsoever, a counterfactual question about that simulation. For instance, suppose our world were a simulation and you were to ask its programmer what would happen if I hit this computer with a sledgehammer. The very first question out of the programmer's mouth would be, "What kind of simulation are we talking about? What's its programming like? Was there a *law* programmed into its code to ensure that the future happens just like the past? If so, then yes, if X were to hit his computer with a sledgehammer, then the computer would be smashed. However, if not—if the simulation is programmed such that all previous sledgehammer-hittings are followed by smashing-effects but *this* sledgehammer hitting was programmed to cause a big surprise (i.e. a non-smashed computer), then the same counterfactual would be false." In other words, although it is entirely natural to think that we can know which counterfactuals are true on the basis of our experiences of regularities in our world (e.g. electrocutions following electric-wire touchings, sledgehammer hittings followed by smashings, etc.), *in a simulation this is actually false*: one cannot know which counterfactuals are true of any given simulation "from the inside"; one has to know *what kind* of simulation one is in (i.e. how it is actually programmed). As surprising as this might seem, I submit, it is actually not the least bit controversial. Anyone who knows anything about programming would tell you precisely the same thing: what a program *counterfactually would do* depends entirely on *how it is actually programmed*. The strange implication here—that we cannot know which counterfactuals are true in our world, if our world is a simulation—is simply a commonsensical implication of a "strange" (but nevertheless epistemically and metaphysically possible) hypothesis: the hypothesis that our reality is a computer simulation. Once one takes the simulation hypothesis seriously, one can see that the strange implication (i.e. one has to know a simulated world's programming *before* one can know which counterfactuals are true of it) is not strange at all.

Objection #5-changing the subject?: A fourth and final objection is that my argument is a kind of red herring. "Sure", it might be conceded, "you have given some good reasons for believing that the Lewis-Stalnaker analysis fails—and that we cannot know which counterfactuals are true—if we are in a simulation. But this is not really a refutation of the Lewis-Stalnaker analysis. It is at most a conditional refutation of sorts: an argument that *if* we are in a simulation, then the Lewis-Stalnaker analysis is false. But this should not worry proponents of the Lewis-Stalnaker analysis too much. I'm happy to admit that if we live in a simulation, all bets are off. Maybe then we do not know which counterfactuals are true, etc. Still, it seems exceedingly unlikely that we are in a simulation, and if we are not, the Lewis-Stalnaker analysis still stands."

Reply: This objection, while tempting, is misguided for several reasons. First, either the Lewis-Stalnaker analysis is correct, or it isn't. If it is correct, however, then purely as a matter of logical consistency it must apply to simulated worlds. For simulated worlds, whatever else they are, are parts of worlds (e.g. if our "world" is a computer simulation, then it—the computer system—exists within a larger, concrete "meta-world"). Since simulations are parts of worlds, for the Lewis-Stalnaker analysis to be correct at all, it has to give the right kinds of truth-conditions for those parts of the world (i.e. the simulation). But, as we have seen, this is false. The Lewis-Stalnaker analysis gives the wrong truth conditions for simulations, both Humean simulations and Necessitarian ones. Second, even if it weren't incoherent in this way, the objection would "give the game away." It would amount to a concession that the Lewis-Stalnaker analysis is false if our world is a simulated one. But, in that case—even if one thinks it is unlikely that our world is a simulation—it is still the case that the Lewis-Stalnaker analysis is false for a significant array of possible worlds. Indeed, the objection concedes that we have to know whether we live in a simulation "in order to know whether the Lewis-Stalnaker analysis is correct." But this in itself is a refutation of the theory. For the theory has never been taken to depend on what kind of world we live in. It has been proposed as a theory of the semantics of counterfactuals simpliciter. Finally, I believe, the arguments given earlier show—very distinctly—why the Lewis-Stalnaker analysis was never well-motivated to begin with. For let us recall the reasons why a programmer would not be tempted to analyze counterfactuals for a Humean simulation in terms of "nearby simulations." The reason for this was simple: what happens in any given Humean simulation in no way depends on any other. The very same thing, however, is true if our world is Humean and not a simulation. If our world is Humean, then what happens in our world depends in no way on what may or may not happen in any "nearby possible world." If our world is Humean and I touch a wire and get a shock, then the counterfactual, "If I touched the wire, I would get shocked", is true. However, if our world is Humean, then—just as with a Humean simulation—we cannot state whether the counterfactual, "If I had not touched the wire, I would not have been shocked", until we *stipulate* which counterfactual reality we are interested in (viz. "If our world had been such that Jones never touched the wire and did not get shocked, then it would be true that if he hadn't touched it, he wouldn't be shocked. However, if our world had been such that Jones never touched the wire but felt a shock anyway, *then* if he hadn't touched the wire, he *would* have been shocked"). Conversely, if our world is Necessitarian—if it is governed by *laws*—then, for the very same reasons as a Necessitarian simulation, the Lewis-Stalnaker analysis is once again false: albeit for different reasons—namely, that counterfactuals in a Necessitarian reality are given by that reality's laws and initial conditions.

§4. Skeptical (But True) Implications?

As noted earlier, if my argument is sound—as I believe it to be—it has very some striking implications above and beyond the falsity of the Lewis-Stalnaker analysis. The first striking implication is this: we cannot possibly know which counterfactuals are true in our world without knowing which kind of world (Humean, Necessitarian, or otherwise) we reside in above and beyond those that are trivially true (viz. if I touch a wire and am shocked, it is trivially true at my world that if I were to touch it, I would be shocked). The reasons for this, again, are simple: if I am right, the semantics for counterfactuals depends on which kind of world we are in. On the one hand, if we are in a Humean world, then each possible world's actual series of events constitutes which counterfactuals are true of that world, and other possible worlds express alternative counterfactual relationships. On the other hand, if we are in a Necessitarian world, then the semantics for counterfactuals are given by a world's governing laws and possible initial conditions (viz. if our world is governed by a Necessitarian law of gravity, it's true that if I were dropped off a ledge, gravity would accelerate me toward the ground—something which is true not because of "nearby possible worlds", but simply because this world has those laws and could have had different initial conditions). All of which brings us to a second, even more stupefying implication: namely, that we cannot know from within our world which semantics for counterfactuals is actually true. For, as noted above, a Humean reality and a Necessitarian reality might look identical from the inside (i.e. from the standpoint of observers within them). Because different counterfactuals are true for Humean and Necessitarian worlds, the only way to know which semantic analysis is correct is to know which type of world one lives in. But this, again, is impossible from within any such world. One can only know which of world ours is from the outside—e.g. from the perspective of a programmer (or Creator).

Is all of this absurd? Are these implications too much to stomach? Philosophers have something of an unfortunate habit of rejecting arguments that have implications they don't like. "That's counterintuitive", is, we all know, a common enough refrain—a refrain invoked for the sake of rejecting a given theory or argument. Yet it is not, I submit, the philosopher's right to make this kind of move. Philosophy should be concerned not with what "seems intuitive." It should be concerned instead with what is true—and, as quantum physics and relativity have shown, reality is under no obligation to conform to our "intuitions." If this paper's argument is sound—as I believe it to be—it's implications are true. We cannot know which counterfactuals are true of our world from our position within it. We cannot know which theory of the semantics of counterfactuals is true either. The *reality* is this: whatever counterfactual dependencies our world instantiates depends entirely on what kind of world ours is at an ultimate metaphysical level—a Kantian "noumenal" level that can, in principle, only be apprehended from outside of our world, from the perspective of a programmer or Creator, not within.

References

Arvan, Marcus (2014). "A Unified Explanation of Quantum Phenomena? The Case for the Peer-to-Peer Simulation Hypothesis as an Interdisciplinary Research Program." *The Philosophical Forum*, 45(4): 433-446.

----- (2013). "A New Theory of Free Will." *The Philosophical Forum*, 44(1): 1-48.

Beane, Silas; Zohreh Davoudi, Martin J. Savage (2012). "Constraints on the Universe as a Numerical Simulation." arXiv:1210.1847.

Bennett, Jonathan (2003). *A Philosophical Guide to Conditionals* (Oxford: Oxford University Press).

Bostrom, By Nick (2003). Are we living in a computer simulation? *Philosophical Quarterly* 53 (211):243–255.

Chalmers, David J. (2005). The matrix as metaphysics. In Christopher Grau (ed.), *Philosophers Explore the Matrix*. Oxford University Press. 132

Jacobs, Jonathan D. (2010). A powers theory of modality: or, how I learned to stop worrying and reject possible worlds. *Philosophical Studies* 151 (2):227-248.

Kripke, Saul (1972). "Naming and Necessity." In Donald Davidson and Gilbert Harman (eds.), Semantics of Natural Language, 253–355, 763–9. Dordrecht: D. Reidel. Revised edition published in 1980 as Naming and Necessity (Cambridge, MA: Harvard University Press)

Lewis, David K. (1986). *On the plurality of worlds* (Vol. 322). Oxford: Blackwell.

----- (1979), “Counterfactual Dependence and Time's Arrow,” *Noûs*, 13: 455–476.

----- (1973). *Counterfactuals* (Oxford: Blackwell Publishers and Cambridge, MA: Harvard University Press, Reprinted with revisions, 1986).

Sider, T. (2006). Beyond the Humphrey objection. *Unpublished manuscript*.

Stalnaker, Robert (1968). “A Theory of Conditionals, ” in Nicholas Rescher (ed.), *Studies in Logical Theory* (American Philosophical Quarterly Monograph Series: Volume 2), Oxford: Blackwell, pp. 98–112.

Weatherson, Brian (2014). "David Lewis", *The Stanford Encyclopedia of Philosophy* (Winter 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2014/entries/david-lewis/>.