Counterfactuals and their Truthmakers: Comparing the Relative Strengths and Weaknesses of Plato and Lewis

Draft; Do Not Cite
Final Version: *Polish Journal of Philosophy* 8.2 (2014): 7-24
DOI: 10.5840/pjphil20082219

Abstract. This article compares David Lewis's understanding of counterfactuals with a Platonic theory of counterfactual truthmakers. By pointing to some weaknesses in Lewis's theory, it will highlight some of the strengths of the Platonic theory. The article will progress in the following way. First, I present David Lewis's understanding of counterfactuals, and discuss some problems the theory has. Next, I discuss Platonic truthmakers, in general, and then show how this applies to counterfactuals. Finally, I discuss the strengths and weaknesses of the Platonic theory, and how it is superior to Lewis's theory.

Counterfactuals are notoriously ill-behaved things. Take, for example, "If Tom had made that catch, we would have won the game;" it seems that most everyone knows what is meant by the statement.[1] However, whether or not the statement is true, or what it is that makes it true, is much more problematic. Here is the issue: despite the fact that it seems that most people know what counterfactuals mean, intuitions and vague understandings

[1] Whether people *actually* know what such statements mean is an entirely different concern, but this just illustrates the complexity of the issue under consideration.

cannot do any theoretical heavy lifting. Thus, any theory that depends on counterfactuals – e.g. an understanding of causation, safety conditions for knowledge – is in trouble if it turns out that what makes counterfactuals true, or false, does not divide up the cases in the right way.

The purpose of this article is to explore counterfactual truthmakers. In particular, I will be presenting a Platonic theory of counterfactual truthmakers. As a way into the issue, I will be using David Lewis's possible worlds understanding of counterfactuals as a foil. By pointing to some weaknesses in Lewis's theory, it will highlight some of the strengths of the Platonic theory I will be presenting.

The article will progress in the following way. First, I present David Lewis's understanding of counterfactuals, and discuss some problems the theory has. Next, I discuss Platonic truthmakers, in general, and then show how this applies to counterfactuals. Finally, I discuss the strengths and weaknesses of the Platonic theory, and how it is superior to Lewis's theory.

1. Lewis and Counterfactuals

In this section of the article, I present, briefly, David Lewis's account of counterfactuals. More precisely, I present how Lewis believes one is to understand what makes a counterfactual true or false. I then go on to consider some objections to Lewis's account of counterfactuals, and their truth conditions.

To explain his understanding of counterfactuals, and their truth conditions, it is worth quoting Lewis at length.

> Given any two propositions A and C, we have their *counterfactual* A □→ C: the proposition that if A were true, then C would also be true. The operation □→ is defined by a rule of truth as follows A □→ C is true (at a world *w*) iff either (1) there are no possible A-worlds (in which case A □→ C is *vacuous*), or (2) some Aworld where C holds is closer (to *w*) than is any A-world where C does not hold. [… or more simply] A □→ C is nonvacuously true iff C holds at all the closest Aworlds (Lewis, 1986, p. 164).

While what a counterfactual is seems fairly straightforward, there are two components of the truth conditions that need a bit more explication. First, what Lewis means by "possible world" needs to be clarified. Second, is how "closeness" of worlds is to be understood.

Lewis claims that "[i]t is uncontroversially true that things might be otherwise than they are"–and Tom catching the ball, and us winning the game, is just the type of thing Lewis has in mind as something that might have been otherwise (Lewis, 1976, p. 84). Lewis takes it that this is an existentially quantified statement. In other words, if it is true that things might have been otherwise than they are, and that is an existence claim, then there exists something – some entity – that would make it true that things might have been otherwise. Lewis claims that the only things that could make the existence claims of this sort true are possible worlds.

By possible worlds Lewis has in mind entities like the world in which one finds oneself, but different. For example, a possible world would be just like this world except where Tom (or more accurately Tom's "counterpart", an entity that bears an appropriate similarity relation to Tom in the actual world) caught the ball, and we won the game (or the counterpart of our team) – call this world T. So, the reason why it is true, if it is true, that had Tom caught the ball, we would have won the

game is because world T exists.[2] There are two important things to emphasize at this point. First, Lewis thinks the only way that one can make sense of existentially quantified possibility claims – of which counterfactuals are a species – is by being a realist about possible worlds. The alternatives are all unappealing–viz. taking the possibility claims as primitive, which is just the absence of any theorizing, or taking them as "metalinguistic predicates analyzable in terms of consistency" which is inherently circular in that it presupposes, in some sense, an understanding of possibility (Lewis, 1976, p. 85). Second, in being a realist about possible worlds, Lewis denies that possible worlds can be reduced to something else, e.g. sets of sentences. This is so because Lewis believes that possible worlds are not different in kind from the actual world, and it is hard to believe that the actual world is just a set of sentences. The idea is that

[2] Here and elsewhere in the article, for simplicity, I have implied that the existence of one world makes a would-counterfactual true. As will become clear, for a wouldcounterfactual to be true for Lewis, the state of affairs described by the counterfactual would have to obtain at most, if not all, nearby possible worlds, roughly.

some people have also discussed possible worlds, but were not realists about them. Others have thought of worlds as linguistic entities – sets of maximally consistent atomic sentences or states of affairs, or diagrammed models. Lewis mentions these other accounts of possible worlds in passing in order to underscore the fact that when he claims to be a realist about possible worlds, he means that quite literally.[3]

In his theory of possible worlds, Lewis wants to be clear that the world which is called the actual world is just one possible world among many, perhaps infinitely many, and has no privileged place among all possible worlds. What makes this world actual, or actualized, and all other possible worlds unactualized is simply that one finds oneself there. The possible world where "Tom" makes the catch, is unactualized for Tom,

[3] Certainly there are others that might be realists about possible worlds – Plantiga perhaps – though for quite different reasons. Lewis is just trying to distinguish his view from the "linguistic entity" understanding of possible worlds. My thanks to Scott Berman for drawing my attention to this detail.

but is actual for Tom's ball catching counterpart.[4] In short, for Lewis, the term 'actual' is indexical in just the same type of way as the term 'I' or 'present' is. "[I]t depends for its reference on the circumstances of utterance, to wit the world where the utterance is located" (Lewis, 1976, p. 86).

That, then, is the rough idea of what Lewis has in mind when he speaks of possible worlds. Before moving on to discuss closeness of worlds, it is important to understand that possible worlds are all spatio-temporally distinct. Not spatio-temporally distinct in the way that a table is distinct from the chair next to it, but in the sense that "you can't get there from here". Possible worlds have distinct existences. Thus, in a very literal sense, no world can "access"[5] another world.

---

[4] It is not essential, at this time, to discuss Lewis's counterpart theory. For now, one can just think of Tom's counterpart, as Tom's representative at another world. Part of Lewis's motivation for counterparts is to avoid complications of trans-world identity, which Quine found so problematic.

[5] Not accessibility in the modal-logical sense, but in the sense that I have access to my car, the Himalayas, the Civil War, myself tomorrow, and some galaxy on the other side of the universe.

The other important aspect of Lewis's truth conditions for counterfactuals is the concept of closeness – sometimes also understood in terms of similarity – of worlds. Simply put, a world *w* is closer to some world *u* than another world *v*, just in case, *w* is overall more similar – in the relevant respects – to *u* than *v* is. Lewis acknowledges that the concept of similarity is a somewhat vague notion, but that does not entail that one does not know what it means. After all everyone makes comparisons of overall similarity all the time: "Sarah looks just like that famous movie star." or "Those twins, Bob and Jim, are nothing alike."

What is of note in the two examples just given, and bears importantly on Lewis's understanding of similarity is that despite being vague, the ease of use and understandability is fairly straightforward, and the standards of comparison can move around, and are, somewhat, context sensitive. Thus, unless Sarah is the famous movie star, she clearly does not look *just* like the movie star, it is only that in important ways Sarah and the movie star have a sufficient number of important

characteristics in common–e.g. facial structure, hair color, etc. Similarly, since Bob and Jim are twins they really are quite similar, even down to their DNA, what has happened in the context of describing Bob and Jim is that a few characteristics – e.g. personality traits – are taken as the most important metric of similarity, giving almost no weight to anything else.

If, as Lewis maintains, the truth of counterfactuals depends on the vague concept of similarity, then

> [t]he truth conditions for counterfactuals are fixed only within rough limits; like the relative importances of respects of comparison that underlie the comparative similarity of worlds, they are highly volatile matter varying with every shift of context and interest (Lewis, 1976, p. 92).

Moreover, trying to fix a way to give a precise measure of comparative similarity is hopeless. "Not only would we go wrong by giving a precise analysis of an imprecise concept; our precise concept would not fall within – or even near – the permissible range of variation of the ordinary concept" (Lewis, 1976, p. 95). The problem, of course, is that if the truth conditions for counterfactuals are as volatile as Lewis implies,

then it would seem that there is good reason to reject his theory, prima facie.

Lewis does have an answer to this volatility problem. He thinks that the relevant respects of similarity are fixed by convention. Although the standards of what counts for something being similar to something else vary, the standards that most people use tend to fall mostly within a fairly narrow range, and most people expect others' standards to fall within the same range. "It is natural that we should have vocabulary conventionally reserved for use within that mutually expected range." Thus, if a speaker uses standards that fall outside of the "normal" range, the speaker, according to Lewis, is using a different vocabulary. Take the Bob and Jim example, and imagine that not only do they look and sound alike, but they have very similar personalities, and most people cannot even distinguish the two. Now, if someone were to assert that Bob and Jim were nothing alike because Bob has a blue aura and Jim has a red aura, clearly that person would be taken to be saying something that is, at best, rather bizarre. The reason it is bizarre

is precisely because the standards of comparative similarity being applied fall far outside the normal range.

The upshot for Lewis is that the vagueness of similarity accounts for the relative vagueness of counterfactuals themselves.

> It accounts for the fact that some sensitive counterfactuals are so vague as to be unsuitable for use in serious discourse; that others have definite truth values only when context serves to narrow their range of vagueness; and that many more have quite definite truth values (in worlds of the sort we think we inhabit), insensitive to small shifts in our standards of comparative similarity (Lewis, 1976, p. 94).

In other words, Lewis thinks his account explains why counterfactuals are often poorly understood. The reason they are poorly understood is that their truth conditions are often vague, and often highly context sensitive.

Having briefly presented Lewis's understanding of counterfactuals, I now turn to some problems that his account has. First, Lewis's theory rests on his general metaphysical nominalism: if there is reason to doubt nominalism, then there is reason to doubt his theory of counterfactuals. Worded another

way, if one has independent grounds for rejecting nominalism, then one has a sufficient reason to at least be critical of Lewis's theory of counterfactuals. There are many arguments both in favor and critical of nominalism. Surveying this debate for the best arguments extends beyond the scope of this article. Nominalism may or may not be true, but it is one of "the costs" of his theory, which of course one might be willing to accept. Second, there are problems, or costs, with his possible world account, particularly for comparative similarity.

From the time Lewis published his book *Counterfactuals* people have been skeptical of possible worlds and comparative similarity as ways to ground an understanding of counterfactuals. Kit Fine, in a Critical Notice on Lewis's book in *Mind* presents what has become a quite famous counterexample to Lewis's theory.

> The counterfactual 'if Nixon had pressed the button there would have been a nuclear holocaust' is true or can be imagined to be so. Now suppose that there never will be a nuclear holocaust. Then that counterfactual is, on Lewis's analysis, very likely false. For given any world in which antecedent and consequent are both true it will be easy to imagine a closer world in which the

> antecedent is true and the consequent false. For we need only imagine a change that prevents the holocaust but that does not require such a great divergence from reality (Fine, 1975, p. 452).

The idea is that, at least intuitively, it seems that had President Nixon pressed the button to launch nuclear missiles–at the Soviet Union, say– during the height of the Cold War, this would have led to retaliation by the Soviets and almost certainly to a nuclear holocaust. The problem is any possible world where a nuclear holocaust has occurred is radically different from the actual world–where there has been no nuclear holocaust. Thus, a world where Nixon had pressed the button and there was no nuclear holocaust–for whatever reason–is more similar, and thus closer, to the actual world than any nuclear holocaust world. Remember, that according to Lewis's analysis if there is at least one world where the antecedent is true and the consequent is false closer than any world where both the antecedent and consequent are true, then the counterfactual is false. Therefore, the Nixon counterfactual is false, but that seems like the wrong answer.

Lewis, of course tries to remedy Nixon type counterfactuals by adding in things like holding fixed the laws of nature, or making comparative similarity dependent on things progressing the way one normally expects them to, and so forth. However, that seems at best ad hoc, and at worst question-begging. Lewis's "fixes" seem to imply that what counts as similarity when evaluating counterfactuals is determined by the counterfactuals themselves. But, the similarity of worlds was supposed to explain the truth of counterfactuals, not vice versa. In other words, if the counterfactuals that one takes to be true determine what counts as a relevantly similar world, then similar worlds do very little to explain why and how counterfactuals are true.

Moreover, if Fine is right then,

> the notion of comparative similarity gives rise to an immediate danger of circularity. For similarity is a matter of agreement in propositions; and among those propositions will be counterfactual ones. So to evaluate a counterfactual, one needs to compare worlds for similarity to the actual world and this would seem to require the evaluation of further counterfactuals. […Thus,] it is no longer clear what the truth of a counterfactual consists in or how we can ever come to

> know that counterfactuals are true or false (Fine, 1975, p. 455).

The idea here is that even if Lewis can give an account of how a given counterfactual does not determine what counts as a relevantly similar world, there is still a danger of circularity. It seems that one thing that would make one world relevantly similar to another is that the same counterfactuals are true of both, but then one would need to compare worlds to check those counterfactuals, which would in turn involve counterfactuals. One could interpret this back and forth of counterfactuals and similarity as circular, or as an infinite regress. Either way, it is clearly a problem for Lewis's account.

Despite Fine's concerns, something like Lewis's account has been the dominant view of counterfactual truthmakers. While there is some concern with Lewis's understanding of what a possible world is–viz. a concrete real existent–the similarity of worlds, whatever they might be, is still quite common. Recently, Alan Hajek has argued that most counterfactuals are false, and they are false, in part, because "the connection

between similarity and the truth-conditions for counterfactuals is far less straightforward than has been widely assumed" (Hajek, unpublished, p. 14). His point is that for any ordering for determining similarity of worlds, which are almost always based on intuitions, it is susceptible to counterexamples that lead to unintuitive results – e.g. Fine's Nixon counterfactual. However, just based on the semantics of counterfactuals, and the chanciness, or indeterminacy involved, a Lewisian understanding of counterfactuals is problematic.

By definition there is a duality between would- and might-counterfactuals. Would-counterfactuals assert something of the form: if Φ were to obtain, Ψ would obtain–formally, $\Phi \Box\!\rightarrow \Psi$. Might-counterfactuals assert something of the form: if Φ were to obtain, Ψ might obtain–formally, $\Phi \Diamond\!\rightarrow \Psi$. Now, by definition–in fact Lewis's definition–$(\Phi \Box\!\rightarrow \Psi) \leftrightarrow \sim(\Phi \Diamond\!\rightarrow \sim\Psi)$ (Lewis, 1976, p. 21). Less formally, if Φ were to obtain, Ψ would obtain, if and only if it is not the case that if Φ were to obtain, Ψ might not obtain. Now since there is this duality, the truth conditions for the mightcounterfactual are: The operation

◊→ is defined by a rule of truth as follows A ◊→ C is nonvacuously true (at a world *w*) iff C holds at least one Aworld (Lewis, 1976, p. 21).

To make his objection more concrete Hajek considers the claim, "if I were to flip this coin, it would land heads." Now that seems false, and the reason it seems false is because if I were to flip this coin, it might land tails (Hajek, unpublished, 3). What Hajek wants to suggest is that almost all would-counterfactuals are just like the coin flipping example. Thus, the reason the Nixon counterfactual is false is because there might not have been a nuclear holocaust if Nixon had pressed the button.

Ultimately, Hajek is pushing Lewis, and Lewisians, on the comparative similarity issue. Hajek wants to draw attention to the fact that there is no good way to rank, or order, the closeness of worlds. If there is no good way to rank the closeness of worlds, then the only reasonable thing to do is consider probabilities. So, the reason the coin flip counterfactual is false is because there is a greater than zero probability – actually a

probability of approximately .5 – that the coin will land tails. The reason the Nixon counterfactual is false is because there is a greater than zero probability that a nuclear holocaust would not have occurred.

Part of Hajek's motivation is that the current best science–quantum mechanics, for example–maintains that the laws of nature as we know them are indeterministic and probabilistic. "And it isn't just the canonical quantum mechanical examples – radioactive decay, spin measurements on a particle in a Stern-Gerlach apparatus, and so on – that are indeterministic. The indeterminism reaches medium-sized dry goods (and even oversized wet ones), just less obviously so" (Hajek, unpublished, p. 7). Hajek gives a billiard ball example to drive the point home.

> Two billiard balls colliding may approximate a deterministic system, but even they are not immune from quantum mechanical indeterminism. One ball might spontaneously tunnel through the other, to China, or to the North Star–incredibly unlikely, to be sure, but possible. Thus I cannot truly say "if the cue ball were to hit the 8 ball, the 8 ball would begin rolling" (Hajek, unpublished, p. 7).

However, it should be noted that even determinism would not eliminate the chanciness that is required to make the might-counterfactual true. Hajek points out that a prime example occurs in statistical mechanics–a deterministic system–with Maxwell's demon,[6] but "[t]he point generalizes to other deterministic systems. For every set of initial conditions in which the cue ball hits the 8 ball and each follows an expected trajectory, there is a nearby initial condition in which the balls behave anomalously" (Hajek, unpublished, p. 20).

Ultimately, the point is that even if one grants Lewis everything in his account, there are still problems. What Hajek is drawing attention to is that even if there is a rough idea – granting Lewis his conventionalism regarding standards of similarity – of what counts as a similar world, there is always the possibility that there is a world, including the actual world where things behave unexpectedly. However, based on the

[6] Hajek (unpublished, p. 20). It seems unimportant to go into the details of what Maxwell's demon is, since the generalized point to follow should be clear enough.

duality of the would- and might-counterfactuals all that there needs to be is a single possible world for a might-counterfactual to be true, then the corresponding wouldcounterfactual is false.

Setting aside, at least in part, the comparative similarity issue there seem to be some problems with possible worlds themselves. In order not to fall into the quagmire of individuation that Quine fears, Lewis denies transworld identity. So, because there is Tom's ball catching counterpart–call him Tom-W–at a world where the counterpart we, the team, win the game– call this world W–it is true that if Tom had caught the ball we would have won the game.[7] Now, since there is not trans-world identity, the fact that Tom-W caught the ball seems to have nothing to do with the actual we, possibly winning our actual game. Lewis talks about "ties of resemblance" and "counterparts", but it is unclear how that is supposed to help. Tom has "ties of resemblance" and "counterparts" at the actual world. One has seen other game winning catches; Tom himself

[7] See note 2.

may have previously made game winning catches. In fact, part of the reason one knows what it means that if Tom had made the catch, we would have won the game is that others–i.e. Tom's counterparts–had, in fact, made game winning catches, and Tom has ties of resemblance to them, including possibly himself. If these type of counterparts help provide the meaning of the counterfactual claim, it seems that they could be potential truthmakers for the claim. After all, one has "access" to these counterparts; one has no access to counterparts at unactualized worlds.[8]

Before concluding this section of the article, it is important to return to Lewis's conventionalism regarding the standards of similarity. His account is plausible enough, and might be a decent explanation of the phenomena of comparative similarity. What people mean, and conventions of discourse, might be a good starting place, but grounding truth on it seems, at least potentially, problematic. It once was convention that there was

[8] See note 4.

some sort of necessary connection between race and intelligence, however that is just false. It is convention that, in the United States, people drive on the right hand side of the road, and if one does not, there are potentially severe consequences, but that does not mean that it is metaphysically true that driving on the right hand side of the road is correct.

Moreover, if the truthmakers, possible worlds, are going to be understood and to function in such a way as to depend on the conventions of normal discourse, Lewis could have simplified his theory greatly. There are already conventions of normal discourse that ground the use and understanding of counterfactuals, in fact, that is why there is an intuitive grasp of their meanings that lead to Fine's and other's counterexamples. Thus, there seems little motivation for adopting Lewis's account, since what he adds just complicates the everyday understanding of counterfactuals, only to end up with an equally questionable everyday understanding of comparative similarity.

2. Plato and Counterfactuals

In this section of the article I present and discuss a Platonic understanding of counterfactuals. I begin by discussing Platonic truthmakers, generally, and then show how the general theory applies to counterfactuals. Anyone familiar with Platonic metaphysics already knows that the Forms function as the truthmakers for Plato–either directly, or indirectly. A complete account and defense of the Forms, while interesting, extends beyond the scope of this article. All that is necessary is to show how the Forms function as truthmakers. A good way to demonstrate how Forms function as truthmakers is to look at the allegory of the cave in the *Republic*. To begin, then, I quote Plato at length.

Imagine human beings living in an underground, cavelike dwelling[. …] They've been there since childhood, fixed in the same place, with their necks and legs fettered able to see only in front of them[. …] Light is provided by a fire burning far above and behind them. Also behind them, but on higher ground, there is a path stretching between them and the fire. Imagine that

along this path a low wall has been built, like the screen in front of puppeteers above which they show their puppets. […]
Then also imagine that there are people along the wall, carrying all kinds of artifacts that project above it–statues of people and other animals, made out of stone, wood, and every material. And, as you'd expect some of the carriers are talking, and some are silent. […]

And if they [the prisoners] could talk to one another, don't you think they'd suppose that the names they used applied to the things they see passing before them? […]

And what if their prison also had an echo from the wall facing them? Don't you think they'd believe that the shadows passing in front of them were talking whenever one of the carriers passing along the wall was doing so? […]

Then the prisoners would in every way believe that the truth is nothing other than the shadows of those artifacts (514a-515c) (Plato, 1997b, pp. 1132-3).

The idea is that the prisoners take to be the real objects of knowledge shadows on the cave wall cast by the artifacts

passing in front of the fire. So, imagine that someone is carrying the statue of a human in front of the fire. The prisoners will take the shadow of the statue to be a human. Now what makes it true that there is what the prisoners take to be a human, is that there is, in fact, the statue of a human passing before the fire, and casting a shadow on the cave wall. Thus, the truthmaker for the shadow human is the statue of the human.

It must be remembered that the allegory of the cave is an allegory. Plato maintains it is an allegory of the human condition. As Julia Annas states: "[t]he prisoners are 'like us', says Socrates (515a). The Cave is, then, not just the degraded state of a bad society. It is the human condition" (Annas, 1997, p. 153). By analogy, what one takes to be a human is the equivalent of the shadow on the cave wall, and the statue of the human passing before the fire is the Form of humanness. Thus, when one perceives a human, what makes it true that there is a human is the Form of humanness, casting its shadow–i.e. participating, or instantiating itself–on the "cave wall" of space-time.

As just described, the connection between the Forms/artifacts and perceptibles/shadows is fairly straightforward. One Form or universal instantiated in one perceptible. However, the process is much more complex. In the cave, the statue is a particular height, shape and texture, and it creates a unique kind of shadow on the cave wall. Likewise in the actual world. It is not just the Form of humanness, but also the Form of a particular length, position and so forth instantiated in a particular area of space-time. Thus, for almost any perceptible in space-time there are myriad Forms participating together that make the perceptible what it is. So, the truthmaker for perceptibles is the complex instantiation of Forms.

So, if it is the Forms that function as truthmakers, then for any counterfactual the Forms must be what makes that counterfactual true. Returning to the cave, how the forms function as truthmakers for counterfactuals can be demonstrated.

> And if there had been any honors, praises, or prizes among them for the one who was sharpest at identifying the shadows as they passed by and who best

> remembered which usually came earlier, which later, and which simultaneously, and who could thus best divine the future (516d) (Plato, 1997b, p. 1134).

Imagine that it just so happens that whenever the statue of a horse passes before the fire, it is followed by the statue of a human. Thus, there is the cave world counterfactual: If a horse were to pass by, then there would be a human not far behind. Now, such a counterfactual is true because had there been a horse-shadow on the cave wall, there would have been a human shadow on the wall shortly after. Moreover, the reason that it is true that the various shadows are cast on the wall is because the various statues pass before the fire.

Now, in the actual world what would make a counterfactual true is the various instantiations of Forms in space-time. So, the Tom-ball-catching counterfactual is true just in case there is a complex relation of Forms that would instantiate in such a way that it is true that Tom catches the ball, and we win the game. More precisely, the Forms must instantiate in such a way that whenever certain Forms come together to make it true that Tom catches the ball at a particular area of space-time, it entails that

other Forms come together to make it true that we win the game at a nearby area of spacetime.

Generally, then, from a Platonic perspective, the truth conditions for counterfactuals are as follows: Given any two states of affairs Φ and Ψ, we have their counterfactual $\Phi \mathrel{\square\!\!\rightarrow} \Psi$: the proposition that if state of affairs Φ were to obtain, then state of affairs Ψ would also obtain. The operation $\mathrel{\square\!\!\rightarrow}$ is defined by a rule of truth as follows $\Phi \mathrel{\square\!\!\rightarrow} \Psi$ iff the complex relation of Forms that make it the case that state of affairs Φ obtains entails, with some type of necessity, that state of affairs Ψ obtains.

Applying these truth conditions to a simple and straightforward case, one gets the following true counterfactual: If Tom's body were in such and such a position, then Tom would be standing. The reason that the Tom standing counterfactual is true is because various Forms must come together to make it true that Tom is in such and such a position – e.g. humanness, straight-leggedness, uprightness, and so-forth – and are instantiated at a particular area of space-time.

However, if all those Forms are instantiated at a particular area of space-time, then, necessarily, the Form of standingness must also be instantiated in that same area of spacetime.

What ultimately makes any counterfactual true is not just the Forms, but the complex interrelations of the Forms. What precisely these complex relations amount to is not necessarily clear, however there are some obvious ones. For example, standingness always goes with straight-leggedness; threeness can never be instantiated with evenness, but is always instantiated with oddness.[9]

Not only do the complex relations of the Forms explain the would-counterfactual, but they also explain the might-counterfactual. Given any two states of affairs Φ and Ψ, we have their might-counterfactual $\Phi \diamond\!\!\rightarrow \Psi$: the proposition that if state of affairs Φ were to obtain, then state of affairs Ψ might also obtain. The operation $\diamond\!\!\rightarrow$ is defined by a rule of truth as follows $\Phi \diamond\!\!\rightarrow \Psi$ iff the complex relation of Forms that make it

[9] For example, Plato (1997a, p. 89).

the case that state of affairs Φ obtains allows for the possibility that state of affairs Ψ obtains.

So, the following might-counterfactual is true: If Tom were not standing, then he might be sitting. The reason that this counterfactual is true is because the complex relation of Forms that make it true that Tom is not standing – e.g. humanness, otherness with respect to standing, and so forth – allow for the possibility that the Form of sitting could be true of the same area of space-time. On the other hand, the counterfactual, if Tom were not standing, then he might be flying, is false. The reason that the flying counterfactual is false is because the Form of flyingness cannot be instantiated in the same area of space-time as the Form of humanness. I use this as an example, and it may not in fact be true, however, there certainly are incompatible forms – e.g. oddness and evenness cannot be coinstantiated – but whatever the example, I think the most natural way to read Plato is that it is a metaphysical impossibility not a nomological one. I am offering the Platonic account as somewhat distinct from Armstrong's and others'

accounts of natural laws. Nomological necessity follows the Forms, not the other way around.[10]

It is assumed that the would-might duality still holds for the Platonic account of counterfactuals. Thus, $(\Phi \mathrel{\Box\!\!\rightarrow} \Psi) \leftrightarrow \sim(\Phi \mathrel{\Diamond\!\!\rightarrow} \sim\Psi)$, and also $(\Phi \mathrel{\Diamond\!\!\rightarrow} \Psi) \leftrightarrow \sim(\Phi \mathrel{\Box\!\!\rightarrow} \sim\Psi)$. So, for example, if it is false that if Tom were not standing, then he might be flying, then it is true that if Tom were not standing, then he would not be flying.

3. Plato v. Lewis

Now there is an understanding of Platonic truthmakers for counterfactuals, and of the truth conditions for both the would- and might-counterfactuals on the Platonic account. In this section of the article I consider some possible weaknesses of this Platonic account of counterfactuals. Despite the fact that

[10] One can think of this as a kind of "Prolegomena" for a Platonic theory of counterfactuals. By giving reasons to think that, at least in some respects, a Platonic theory is superior to a Lewisian one, gives some reason to further develop a full theory that spells out all the implications.

there are some problems with the account put forward in the previous section, it is still superior to Lewis's account, discussed in the first section.

Above, two main objections to Lewis's account of counterfactuals were considered. First, it was pointed out that how Lewis understands counterfactuals depends on his broad metaphysical commitments. Thus, it was suggested that if there is reason to doubt nominalism, then that is sufficient to doubt that his account of counterfactuals is adequate. Second, based on arguments from Alan Hajek and others, it was shown that Lewis's understanding of counterfactuals leads to some unintuitive results, and/or results in the fact that most counterfactuals – at least would-counterfactuals – turn out to be false. I now turn to how the Platonic account of counterfactuals handles similar objections.

As has been mentioned, Lewis's account of counterfactuals is dependent on Lewis's general metaphysics, namely, nominalism. In the same way, the Platonic account of counterfactuals is dependent on the general Platonic

metaphysics. Because there is this dependence, if there is reason to doubt the Platonic metaphysics, then there is reason to question the Platonic version of counterfactuals put forth in the previous section. A full defense of a Platonic metaphysics – or something relevantly similar – extends beyond the scope of this article. Prima facie, though, this "cost" for the Platonic account is roughly equivalent to the "cost" of nominalism for the Lewisian.

There were actually two related objections from Hajek – and Fine – regarding Lewis's account of counterfactuals. First, the relationship between similarity of worlds and counterfactual truth conditions is problematic. Second, based on the would-might duality, it turns out that most counterfactuals – at least would-counterfactuals – turn out to be false. Since the Platonic account put forward above does not depend on possible worlds or comparative similarity, the issues that arise due to the unintuitive results of comparative similarity are not a problem. However, the wouldmight duality still seems to get some traction.

It was granted that the would-might duality holds for the Platonic account of counterfactuals. A more complete account might suggest a different way of understanding the relationship between would- and mightcounterfactuals. However, even if the would-might duality holds, the Platonic version still fares better than the Lewisian.

On the Lewisian account it turns out that not only are most wouldcounterfactuals false, but virtually all of them are. The exceptions would be tautologies, and necessary, logical truths, etc. For example, $\Phi \mathrel{\Box\!\!\rightarrow} \Phi$ will be true for Lewis, also, if it is true that $\Box(P \rightarrow Q)$, then $P \mathrel{\Box\!\!\rightarrow} Q$ would be true. For the Platonic account, those few would-counterfactuals that are true for Lewis are also true. On the assumption that quantum indeterminism holds and the would-might duality also holds, many would counterfactuals, perhaps most, will also be false on the Platonic account. Thus, if I were to flip this coin, it would land heads is false on the Platonic account for the same type of reason that it is false for Lewis–because it is true that if I were to flip this coin, it might land tails. Likewise,

> Two billiard balls colliding may approximate a deterministic system, but even they are not immune from quantum mechanical indeterminism. One ball might spontaneously tunnel through the other, to China, or to the North Star–incredibly unlikely, to be sure but possible. Thus I cannot truly say "if the cue ball were to hit the 8 ball, the 8 ball would begin rolling" (Hajek, unpublished, p. 7).

This billiard ball scenario holds for both the Lewisian and the Platonic accounts of counterfactuals.

Even though it turns out that most would-counterfactuals come out to be false whether the truth conditions are Lewisian or Platonic, this may just be an odd quirk of the semantics of counterfactuals. However, the Platonic account still fairs better than the Lewisian. The reason is that because the Forms have unique and complex ways of interacting, certain states of affairs that are possible in Lewis's possible worlds metaphysics are just not possible on the Platonic account,. Take the counterfactual mentioned above, if Tom were not standing, then he might be flying: it is false on the Platonic account. Because of the way the Forms are, and how they can coinstantiate, it just is not possible that the Form of humanness can be instantiated in the

same area of space-time with the Form of flyingness.[11] Because of the would-might duality, on the Platonic account there is the corresponding true would-counterfactual–viz. if Tom were not standing, then he would not be flying.

For Lewis, the only limit on possibility is the so-called logical possibility. From the Lewisian perspective, it is logically possible that humans could fly. Thus, it is true that if Tom were not standing, then he might be flying, and the corresponding would-counterfactual, if Tom were not standing, then he would not be flying, is false. These types of cases can be multiplied. So, depending on what is true of the Forms and their complex interrelations, various might-counterfactuals that would be true for Lewis will turn out false, but then the corresponding would-counterfactuals that are false for Lewis will be true on the Platonic account. This, then, is something that counts in favor of

[11] As said above, I use this by way of example. The exact states of affairs that are metaphysically impossible would be determined by what is in fact true of the forms – but it is reasonable to assume that this is a much larger class than those which are impossible for the Lewisian.

the Platonic account of counterfactuals over the Lewisian account. In other words, even though most counterfactuals are false from both the Lewisian and the Platonic perspectives, there are more true counterfactuals – would-counterfactuals, anyhow – on the Platonic account.

It has been shown that some of the problems that arise for the Lewisian theory of counterfactuals parallel the problems with the Platonic theory of counterfactuals put forward in this article. Despite these problems, it was also shown that the Platonic account still fares better the Lewisian one. Now, it will be shown that there is one other issue that tilts the scales in favor of the Platonic account. Even though this article is concerned with the metaphysics of counterfactuals–with what in fact makes them true or false, or what is the better account of counterfactual truthmakers–there is an epistemic worry.

It would be nice to *know* whether a counterfactual is true. As will be shown, the Platonic version has a better answer to this epistemic worry. Again, for the purposes of this article, the only counterfactual theories that are on the table are the Lewisian

and the Platonic ones, there are of course other theories, but those are not of concern here. If the Platonic theory of counterfactual truthmakers can be shown to be superior to Lewis's, then a great deal has been shown, even if there might be other ways to think about the metaphysics of counterfactuals.

Hajek's and others' concerns aside, a particular counterfactual is true if the state of affairs described obtains at some–more precisely, at most if not all of the closest–possible worlds. On Lewis's account, if Tom were not standing, then he would be flying is true just in the case where at all the closest worlds where Tom – more exactly, Tom's counterpart – is not standing, Tom's counterpart is flying. In order to *know* if the Tom-flying counterfactual is true, one must *know* whether Tom's counterparts are flying at the worlds where Tom's counterparts are not standing. Here is the issue; the denizens of any possible world are confined to that world. Thus, since

someone at the actual world does not have access,[12] there is a very real sense that one cannot *know* what is going on at any possible world other than the actual world.

To clarify, at the actual world, one would know whether Tom was flying when he was not standing by perceiving Tom. The thing is, someone at the actual world cannot perceive what is occurring at any other possible world. Therefore, it does not seem that one could know whether Tom's counterparts are flying at all the closest possible worlds where Tom's counterpart is not standing. If one cannot know what Tom's counterparts are doing at any possible worlds, *eo ipso* one cannot not know whether the counterfactual "If Tom were not standing, then he would be flying" is true or not.

Setting aside any metaphysical issues, there is a huge epistemic worry for Lewis. It seems that even if Lewis has the metaphysics right, there is no way to know the truth regarding

[12] Again, here I am using access in the everyday sense of the word, not in the modal-logical sense of a delimiter of possibility.

any counterfactual. What is more, one cannot even know the truth regarding might-counterfactuals. So, if Tom were not standing, then he might be flying is true just in case there is some possible world where Tom's counterpart is not standing and he is flying. But again, there seems no way to know what is going on at any possible world. Therefore, even these weaker might-counterfactuals are unknowable for Lewis.

There is one way that the truth of at least the might-counterfactuals could be known for Lewis. So, if at the actual world one knew that Tom had the ability to fly, then one could know that if Tom were not standing, then he might be flying.[13] The problem is that this is not an option open to Lewis, as a nominalist. Further, even if Lewis's nominalism did not block this approach to the knowability of counterfactuals, it would render his possible worlds metaphysics completely useless and unnecessary. The appeal, if there is an appeal, of the possible

[13] This issue of knowing abilities and properties as a way to know the truth regarding possibility claims – of which the might-counterfactual is a species – was brought to my attention by Scott Berman.

world metaphysics was supposed to be its explanatory power. If what makes the truth of a counterfactual knowable are properties, or abilities, or dispositions that are knowable at the actual world, then possible worlds have virtually no explanatory power.

One of the nice things about the Platonic theory of counterfactuals is that it has its epistemology built right into the theory. Plato already has an account of how one can know anything. What it ultimately amounts to is that one knows something about the perceptible world only if one knows the Forms involved. Carrying that over to the knowledge of the truth regarding counterfactuals, one knows whether a counterfactual is true or not–at least with might-counterfactuals – if one knows the Forms involved in the state of affairs described by the counterfactual, and if one knows the truth regarding some might-counterfactual then one knows the truth regarding the would-counterfactual that is the might-counterfactual's dual. So, if one knows the Form of humanness and knows that the Form of flying cannot be co-instantiated in

the same area of space-time as the Form of humanness, then one knows that it is false that if Tom were not standing, then he might be flying. Further, because of the would-might duality one also knows that it is true that if Tom were not standing, then he would not be flying. Interestingly, the type of knowledge that is built into the Platonic account of counterfactuals is precisely the type of knowledge that would be necessary for the Lewisian to know regarding counterfactuals.

Of course, one can question whether or not Plato's account of knowledge is correct, but that is beside the point. What is significant is that Lewis does not have an epistemology available to explain how one can know the truth or falsity of counterfactuals, and the Platonic account does. Therefore, this is another place where the Platonic theory of counterfactuals seems superior to Lewis's theory.

4. Conclusion

This article has been an investigation of two theories of counterfactuals and their truthmakers. First, there was a

discussion of David Lewis's influential theory. After presenting his theory, some of its problems were pointed out. Then there was a discussion of Platonic truthmakers in general, and of how they could be applied to a theory of counterfactuals. Finally, it was shown that despite some weaknesses with the theory, it is still superior to the Lewisian account in the various respects considered.

Once again, it has not been shown that the Platonic account presented here is the right theory of counterfactual truthmakers. In order to do that, it would have to be proven that the Platonic theory is superior to every theory of counterfactuals and their truthmakers. Such a project would be too great for an article of this length. Nevertheless, in proving that the Platonic account has distinct advantages over that of Lewis, an important step in the right direction has been made.

References

Annas, J. (1997). Understanding the Good. In Kraut, R. (Ed.) *Plato's Republic: Critical Essays* (pp. 143-168). New York: Rowman &

Littlefield Publishers, Inc.

Berman, S. (1996). Plato's Explanation of False Belief in the *Sophist*. *Aperion 29* (1), 19-46.

Fine, K. (1975). Critical Notice: *Counterfactuals*. *Mind 84* (335), 451-458.

Hajek, A. (unpublished). Most Ordinary Counterfactuals are (probably) False. http://philrsss.anu.edu.au/people-defaults/alanh/papers/MCF.pdf, accessed November 12, 2011.

Lewis, D. (1976). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Lewis, D. (1986). Causation. In Lewis, D. *Philosophical Papers: Volume II* (pp. 159-72). Oxford: Oxford University Press.

Plato. (1997a). Phaedo. In Cooper, J. M. (Ed.) *Plato: Complete Works* (pp. 49-100). Indianapolis: Hackett Publishing Company.

Plato. (1997b). Republic. In Cooper, J. M. (Ed.) *Plato: Complete Works* (pp. 971-1223). Indianapolis,: Hackett Publishing Company.