NOÛS

NOÛS 52:3 (2018) 530–560
doi: 10.1111/nous.12177

# *Counterpossibles in Science: The Case of Relative Computability**

MATTHIAS JENNY
Massachusetts Institute of Technology

**Abstract**

I develop a theory of counterfactuals about relative computability, i.e. counterfactuals such as

> If the validity problem were algorithmically decidable, then the halting problem would also be algorithmically decidable,

which is true, and

> If the validity problem were algorithmically decidable, then arithmetical truth would also be algorithmically decidable,

which is false. These counterfactuals are *counterpossibles*, i.e. they have metaphysically impossible antecedents. They thus pose a challenge to the orthodoxy about counterfactuals, which would treat them as uniformly true. What's more, I argue that these counterpossibles don't just appear in the periphery of relative computability theory but instead they play an ineliminable role in the development of the theory. Finally, I present and discuss a model theory for these counterfactuals that is a straightforward extension of the familiar comparative similarity models.

## 1. Introduction

It is a well known feature of the orthodox possible-worlds approach to counterfactual conditionals due to Robert Stalnaker (1968) and David Lewis (1973) that

*Thanks to Scott Aaronson, Karen Bennett, Matteo Bianchetti, David Boylan, Ben Burgis, Alex Byrne, Michael Detlefsen, Richard Dietz, Kevin Dorst, Nicole Dular, Hartry Field, Melvin Fitting, Juliet Floyd, Cameron Gibbs, Cosmo Grant, Geoffrey Hellman, Jared Henderson, Harold Hodes, Douglas Jesseph, Justin Khoo, Hilary Kornblith, Teresa Kouri, Rose Lenehan, Matt Mandelkern, William Nalls, Eileen Nutting, Alex Paseau, Milo Phillips-Brown, Graham Priest, Agustín Rayo, Bernhard Salow, Chris Scambler, Melissa Schumacher, Kieran Setiya, Stewart Shapiro, Jeremy Shipley, Bradford Skow, Theodore Sider, Robert Stalnaker, Stephen Yablo, Elena Ziliotti, audiences at the 3rd Seoul Philosophy Graduate Student Conference, the 16th Midwest PhilMath Workshop, the 1st Massachusetts-Rhode Island Graduate Workshop in Philosophy, the 2016 Logic and Metaphysics Workshop at the CUNY Graduate Center, the 2016 CSHPM/SCHPM Annual Meeting, and especially Vann McGee and an anonymous referee for *Noûs* for helpful discussion. I also gratefully acknowledge support by the Swiss Study Foundation, whose travel grant allowed me to carry out parts of this project.

it makes all counterfactuals with metaphysically impossible antecedents come out vacuously true. Many have pointed out that this so-called *vacuity thesis* runs counter to our intuitions about the truth-values of many such counterpossibles. Some of the proposed counterexamples to the vacuity thesis concern philosophical questions such as what would be the case if the laws of metaphysics had failed or if certain moral principles had been different, while others are about more ordinary topics such as whether anyone would have cared if Hobbes had squared the circle or what I would do if I were you.[1] Against such proposed counterexamples, Timothy Williamson (2007; 2010; 2015) has recently mounted a fresh defense of the vacuity thesis by making a strong case for its many theoretical virtues. We have thus reached an impasse where theoretical virtues are pitted against intuitive judgments.

In this paper, I put forth a reason for abandoning the vacuity thesis that doesn't rest on our intuitions. I do so by discussing a new source of trouble for the orthodoxy: relative computability theory. Textbook writers often introduce relative computability with the help of counterfactual conditionals. For example, Martin Davis writes that relative computability theory is concerned with the following:

> We may ask, of a given problem $P$,
>
> *If we could solve P, what else could we solve*?
>
> And, we may ask,
>
> *The solutions to which problems would also furnish solutions to P*?
>
> (Davis 1958, 179, emphasis in the original)

After providing some background on relative computability theory, I will argue that, just like other mathematical facts, the facts uncovered by relative computability theorists are metaphysically necessary. So, on the assumption that $P$ is not in fact solvable, the vacuity thesis would have it that any answer to Martin's first question is true.

But the vacuity thesis doesn't just find counterexamples in the way relative computability theorists talk about their discipline in ordinary language. I will argue that non-vacuous counterpossibles play an ineliminable role in relative computability theory. The vacuity thesis thus threatens to declare an established science as nonsensical.

Instead of abandoning relative computability theory in light of this, I will instead draw from its resources to patch up the orthodoxy about counterfactuals. Like previous attempts to revise the theory of counterfactuals, I will present a model theory that makes use of so called 'impossible worlds,' world-like entities where metaphysical impossibilities can hold.[2] However, unlike earlier attempts, which have run into trouble when it comes to expanding Lewis' comparative similarity relation to these new entities,[3] I show that with the right choice of the set of 'worlds,' a comparative similarity relation immediately falls out of the mathematical theory of relative computability that gives the right results for counterfactuals about this theory. Questions remain about how to interpret my proposed model theory, especially the 'worlds' involved. But given the continuity with the comparative

similarity models for ordinary counterfactuals, these questions become tractable. I close by drawing some tentative lessons for the general study of counterpossibles.

## 2. Background

Computability theory studies what sets of natural numbers are algorithmically decidable (or 'solvable,' as in the above quote by Davis). By algorithmic decidability we mean that a computing agent could, in principle, decide for any natural number whether it is a member of the set by mechanically following a completely explicit algorithm that terminates in the right answer in finite time and after finitely many steps. An example of an algorithm is the truth table method, which allows us to decide for any sentence of the propositional calculus whether it is a tautology. The sets of natural numbers whose algorithmic decidability or lack thereof is of particular interest are those that represent certain well-formed problems. *The validity problem* (sometimes simply called *the decision problem*) is the set that encodes the sentences of the predicate calculus that are logically valid.[4] To say that the validity problem is algorithmically decidable would be to say that there is an algorithm that would allow us to decide for any number representing a sentence of the language of the predicate calculus whether it is a member of the set of the validity problem and so whether it is logically valid. It was a significant discovery by Alonzo Church (1936a; 1936b) and Alan Turing (1936) that the validity problem is not algorithmically decidable. Other sets that aren't algorithmically decidable are *the halting problem*, which encodes the problem of deciding whether a computer will eventually halt when it's given a certain input, and *arithmetical truth*, which encodes the true sentences of the language of arithmetic.

As we already saw, *relative* computability theory is introduced by Martin Davis using counterfactuals. Similarly, Hartley Rogers says (where to *calculate* the characteristic function of a set amounts to algorithmically deciding the set):

> Intuitively, $A$ is *reducible* to $B$ if, given any method for calculating [the characteristic function of B], we could then obtain a method for calculating [the characteristic function of A.] (Rogers, Jr. 1967, 127, emphasis in the original)

And, for a more recent example, Herbert Enderton writes:

> On the one hand, we might be able to show that if, hypothetically speaking, we could somehow decide membership in $B$, then we could decide membership in $A$. This would lead us to the opinion that $A$ is no more undecidable than $B$ is. (Enderton 2011, 121)

The study of relative computability was spearheaded by Turing (1939) and Emil Post (1944) and later developed into a mature mathematical discipline using the usual extensional tools of set theory and first-order logic by the likes of Richard Friedberg, Stephen Kleene, Albert Muchnik, Rózsa Péter, and Post.[5] In formal regimentations of mathematics, the only conditional available is of course the material conditional. We know now that counterfactual conditionals behave very differently from material conditionals, but it wasn't until a few years after Turing's and Post's early papers appeared that counterfactual conditionals were identified as interesting

objects of study.[6] And it took another twenty years after that until the now standard possible worlds model theory for counterfactuals was worked out by Robert Stalnaker and David Lewis.[7] But with hindsight, we can ask the kind of questions that we will be presently concerned with.

A basic result of relative computability theory is that the halting problem is *reducible* to the validity problem.[8] This fact can be expressed as follows:

> (valid > halt) If the validity problem were algorithmically decidable, then the halting problem would also be algorithmically decidable.

By contrast, arithmetical truth is *not* reducible to the validity problem. This means that the following is false:

> (valid > arith) If the validity problem were algorithmically decidable, then arithmetical truth would also be algorithmically decidable.

How do we know this? And how, for that matter, do we know that the validity problem, the halting problem, and arithmetical truth aren't algorithmically decidable? We know all of this due to a combination of mathematical theorems and two principles connecting the mathematical apparatus with the notions of algorithmic decidability and reducibility. Take first the fact the validity problem, the halting problem, and arithmetical truth aren't arithmetically decidable. Church and Turing established certain mathematical theorems that get us halfway towards establishing this fact. It will be most illuminating to follow Turing's presentation of the result. Turing introduced a class of abstract machines that are now called *Turing machines*. He then showed that the assumption that, say, the halting problem is decidable by a Turing machine leads to a contradiction, akin to the contradiction Cantor derived from the assumption that the cardinality of the natural numbers is equal to the cardinality of the real numbers. Therefore, there isn't a Turing machine that decides the halting problem, or the validity problem or arithmetical truth for that matter. This is the mathematical part. The other part involves what is nowadays called *the Church-Turing thesis*. This thesis says that the sets that are algorithmically decidable in the informal sense, i.e. the sets that are decidable by any algorithmic means, are just the sets that are decidable by a Turing machine. Interpreted most conservatively, this thesis claims that Turing machines are an adequate model of algorithmic decidability.[9] Putting the Church-Turing thesis together with the fact that there's no Turing machine that decides the halting problem, the validity problem, or arithmetical truth gives us that these sets are algorithmically undecidable.

We know that the halting problem is reducible to the validity problem but arithmetical truth isn't for similar reasons, but with a twist. To establish these results, we need *oracle Turing machines*. An oracle Turing machine is just like a Turing machine, except that it has access to an 'oracle.' Oracles can be thought of as external storage devices that contain the correct answer to any 'yes' or 'no' question about a particular decision problem we may ask them. For example, an oracle for the validity problem contains, for arbitrary sentences of the predicate calculus, the answer to the question whether they are logically valid or not. Think of an oracle Turing machine as just like an ordinary Turing machine, except that it has an extra

port where we can plug in an oracular storage device.[10] We can now show that an oracle Turing machine with an oracle for the validity problem can algorithmically transform the answers it gets about the validity problem into answers about the halting problem. That's how the halting problem is *Turing reducible* to the validity problem. However, even if the oracle Turing machine can ask the oracle questions about the validity problem, it won't be able to transform these answers into answers about arithmetical truth. That's how arithmetical truth isn't Turing reducible to the validity problem. To get from these results, which can be stated and proved purely mathematically, to the results that the halting problem is reducible simpliciter to the validity problem but that arithmetical truth isn't, we need an analogue of the Church-Turing thesis. This thesis, which is variously called *the Post-Turing thesis* or *the relativized Church-Turing thesis*, says that a set $B$ is reducible simpliciter to a set $A$ iff $B$ is Turing reducible to $A$.[11]

But what is this relation of reducibility simpliciter? We may understand the claim that $B$ is reducible to $A$ as saying that if $A$ were algorithmically decidable, then $B$ would be algorithmically decidable—hence the counterfactual locutions in the above quotes from Davis, Rogers, and Enderton. In fact, I will argue that this is *the* way of understanding the claim. This understanding runs into philosophical trouble, however. For it is plausible that facts about what is and isn't algorithmically decidable are metaphysically necessary. The mathematical theorems involved in showing that none of our three sets can be decided by a Turing machine hold of course as a matter of metaphysical necessity. That it's metaphysically necessary that none of the sets are algorithmically decidable then follows by the fact that the Church-Turing thesis is metaphysically necessary.

What reasons do we have for thinking that the Church-Turing thesis is metaphysically necessary? Note that the limits of computation that Church and Turing discovered aren't merely technological. Church and Turing didn't merely show that we haven't built the right kind of computer or discovered the right kind of algorithm to decide the validity problem. In fact, Church and Turing's result *predates* the modern computer. Before anyone had built anything resembling a modern computer, Church and Turing had already identified computational problems that no computer could ever decide. And since the invention of the first computer, all technological innovations in computing, including innovations involving quantum computers that are yet to be realized,[12] have merely lead to an increase in computing speed and efficiency; they never have and never will lead to an improvement in what can be algorithmically decided. Furthermore, the limits of computation that Church and Turning discovered also aren't merely limits imposed by the actual laws of nature. Church and Turing don't argue for their conclusion that the validity problem isn't algorithmically decidable by showing that the laws of nature rule out a computer that decides the validity problem.[13] This suggests that the degree to which it's impossible to algorithmically decide the validity problem is stronger than both technological or nomic impossibility. This suggests, but doesn't yet prove, that the Church-Turing thesis is indeed metaphysically necessary.[14]

There is also a direct argument for the metaphysical impossibility of the claim that, say, the validity problem is algorithmically decidable. To say that the validity

problem isn't algorithmically decidable is just to say that there isn't an algorithm to decide the validity problem. But algorithms are abstract objects that are independent of human activity.[15] As such, they are the kinds of thing that either exist of metaphysical necessity or else don't exist at all; and if they don't exist, then it's metaphysically impossible that they exist. So if there isn't an algorithm to decide the validity problem, then it's metaphysically impossible that there exists such an algorithm, and so it's metaphysically impossible that the validity problem is algorithmically decidable. Thus, (valid > halt) and (valid > arith) are indeed counterpossibles.

The argument just presented relies on certain assumptions about metaphysical possibility and the modal metaphysics of abstracta, assumptions that may be doubted. Nevertheless, the assumptions are perfectly in line with orthodox thinking about these issues. So it follows from orthodox thinking about metaphysical possibility and the modal metaphysics of abstracta that it's metaphysically impossible that the validity problem is algorithmically decidable.[16]

It is important to be clear on what I am and am not claiming. I'm not claiming that it's metaphysically impossible to determine the members of the set corresponding to the validity problem. It is entirely compatible with everything I've said that some deity would be able to tell us for any natural number whether it is a member of that set. But if what I've argued for is right, then even such a deity wouldn't be able to *algorithmically decide* the validity problem, because there is no algorithm that the deity could rely on. But since the antecedent of (valid > halt) and (valid > arith) claims that the validity problem is algorithmically decidable, the metaphysically possible existence of such a deity wouldn't pose a threat to my claim that these counterfactuals are indeed counterpossibles. Of course, given my concession that such a deity may be metaphysically possible, it may be worried immediately that the status of these counterfactuals as counterpossibles aren't significant, for perhaps we can reinterpret counterfactuals about relative computability as about such deities. However, things aren't that simple, as the extended argument in section 4 will show. I'll argue there that such a reinterpretation and many more like it would amount to a revision of what relative computability theorists take themselves to be doing.

But before we move on, let's state precisely what the present challenge to the Stalnaker-Lewis approach to counterfactuals is: we have counterfactuals about relative computability, such as (valid > halt) and (valid > arith) above, some of which appear to be true and some of which appear to be false, but we also have that these counterfactuals have metaphysically impossible antecedents. Now, the usual way of understanding the Stalnaker-Lewis approach to counterfactuals is as follows: a counterfactual ⌜If $\phi$ had been the case, then $\psi$ would have been the case⌝ is true at a metaphysically possible world $w$ iff all metaphysically possible worlds sufficiently similar to $w$ where $\phi$ is true are such that $\psi$ is true in them as well. Since there are no metaphysically possible worlds where the validity problem is algorithmically decidable, any counterfactual that starts with 'If the validity problem were algorithmically decidable . . . ' is vacuously true. Given that with (valid > arith) we have such a counterfactual that appears to be false, we seem to have a counterexample

to the semantics just sketched. And not just that: given that (valid > halt) appears to be true, we also immediately see that we can't just change the orthodoxy so that counterpossibles are all false.[17] And given that the halting problem and arithmetical truth are algorithmically decidable at all the same metaphysically possible worlds—namely none—we also have a counterexample to the part of orthodoxy that says that taking a counterfactual sentence and replacing any of its subsentences with a sentence that's true at all the same metaphysically possible worlds yields a necessarily equivalent sentence.

I said that (valid > arith) *appears* to be false and that (valid > halt) *appears* to be true. In the next two sections, I argue that these appearances aren't deceiving: we ought to understand counterfactuals about relative computability literally; in fact, they play an ineliminable role in the definition of the reducibility relation.

## 3. Philosophical Humility

Researchers in relative computability theory are authorities on the reducibility relation. However, they are generally not experts on the semantics of counterfactuals.[18] So the mere fact that they are disposed to assert some counterfactuals about relative computability and deny others doesn't indefeasibly undermine the orthodoxy about counterfactuals. On the face of it, this fact is simply another piece of evidence that needs to be weighed against the considerations that speak in favor of the orthodoxy, to be filed away with the well-known fact that ordinary speakers are disposed to assert some ordinary counterpossibles and deny others. Perhaps we can hold on to the orthodoxy and excuse relative computability theorists' dispositions by appealing to similar considerations with which we may excuse the dispositions of ordinary speakers. Timothy Williamson (2015, §4), for example, develops an error theory about the dispositions of ordinary speakers. So perhaps we can simply co-opt Williamson's error theory and conclude that counterfactuals about relative computability pose no threat to the orthodoxy, especially in light of the considerable theoretical pressures to hold on to the orthodoxy, also discussed by Williamson (2015, §2).

However, I am going to argue now and in the next section that this response on behalf of the orthodoxy runs counter to a certain kind of philosophical humility. This philosophical humility says that whenever an established mathematical or scientific discipline purports to study a certain phenomenon, we shouldn't give in to philosophical considerations that suggest that there is no such phenomenon to be studied.[19] Relative computability theory, which is certainly an established mathematical discipline, purports to study the reducibility relation. In the previous section, I mentioned that *a* way of understanding the claim that $A$ is reducible to $B$ is as saying that $A$ would be algorithmically decidable if $B$ were algorithmically decidable. I now want to argue that this is *the* way of understanding the reducibility relation. If that's right, and if the orthodoxy about counterfactuals is correct, then the reducibility relation holds between any two algorithmically undecidable sets. It would also mean that $A$ isn't reducible to $B$ iff $B$ is algorithmically decidable and $A$ isn't. But then the reducibility relation would carve out the same distinction

among sets of natural numbers that the property of algorithmic decidability does. The study of the reducibility relation would thus become nothing other than the study of algorithmic decidability, and so relative computability theory is robbed of its own subject matter. In light of this fact, philosophical humility recommends that we reject the vacuity thesis.

Some might argue that philosophical humility should be understood slightly differently. The philosophy of mathematics that emerges from Stephen Yablo's *Aboutness* (2014, esp. §5.3) is a case in point. Astronomers study, among other things, the number of planets. However, nominalists think that numbers don't exist. So nominalism threatens to rob astronomy of one of its subjects. Yablo, who is a nominalist, agrees that astronomers speak falsely when they say that the number of planets in our solar system is eight. However, Yablo thinks that these astronomers nonetheless speak *correctly*, because what they say is partially, and non-vacuously, true—it has a true part, the part that is about the concrete world. Thus, with Yablo's theory of partial truth, we can hold on to a kind of philosophical humility, diminished though it may be, in allowing that that there is a phenomenon that astronomers study: how things stand concretely with the planets. Likewise, perhaps we can extract some non-vacuous core from the claim that $A$ would be algorithmically decidable if $B$ were algorithmically decidable, even when $A$ and $B$ are both algorithmically undecidable. This core would then be the proper phenomenon that relative computability theorists study.

Unfortunately though, Yablo's theory doesn't help in rescuing the orthodoxy about counterfactuals. In order for this theory to yield the result that it's partially true that the number of planets is eight, Yablo needs there to be a possible world where the astronomers' statement is fully true, i.e. a world where numbers exist. Now, think about how we would develop a Yablovian theory of counterfactuals about relative computability. We would say that the statement $(\mathsf{valid} > \mathsf{halt})$ is vacuously true, but its assertion is correct, perhaps because it has a non-vacuously true part that talks about certain structural relationships between the validity problem and the halting problem. But now if we wanted to follow Yablo's theory of partial truth, we would need there to be a possible world where the statement is fully true, and non-vacuously so. In such a world, we would need there to be a possible world where the validity problem is algorithmically decidable. So it looks like our Yablovian theory would require the claim that the validity problem is algorithmically decidable to be possibly possibly true. Now, it's true that unless we help ourselves to the characteristic axiom of the modal logic *S*4, 'possibly, possibly, the validity problem is algorithmically decidable' isn't quite the same as 'possibly, the validity problem is algorithmically decidable.' But it also isn't so far removed from it that we can be said to have made genuine progress on behalf of the orthodox approach to counterfactuals. What's more, both Stalnaker and Lewis as well as the the model theory I will present later validate the *S*4 axiom.

In sum, philosophical humility does indeed recommend that we reject the orthodoxy about counterfactuals, on the assumption that the counterfactual way of understanding the reducibility relation is indeed *the* way of understanding it. I now turn to a defense of this latter claim.

## 4. Understanding and Misunderstanding Reducibility

An immediate reason for thinking that the counterfactual way of understanding the reducibility relation is indeed *the* way of understanding it is that that's exactly how Davis, Rogers, and Enderton characterize the relation in our quotes above (see pages 2–3). But of course, this needn't be decisive. Perhaps we want to say that when relative computability theorists assert (valid > halt), what they're *really* doing is _____. Let's look at seven proposals of how to fill in this blank. The first five are instances of quite general proposals of how to respond to purported counterexamples to the vacuity thesis whereas the final two are specific proposals about our counterfactuals about relative computability. I will argue that none of these proposals work. This suggests that when relative computability theorists assert (valid > halt), they really mean it, which in turn suggests that the reducibility relation is indeed to be understood in terms of counterfactuals.

*Idioms.* Here's a proposal for filling in the blank above: when relative computability theorists assert (valid > halt), what they're *really* doing is assert that the halting problem is reducible to the validity problem; the counterfactual locution (valid > halt) and its variants in the quotes from Davis, Rogers, and Enderton are merely idiomatic ways of gesturing towards the notion of reducibility. Perhaps the counterfactual locution is particularly evocative of some of the ideas behind the notion of reducibility, but sentences such as (valid > arith) aren't to be taken literally.

However, counterfactuals about relative computability don't behave linguistically the way idioms do. In general, idioms, though syntactically complex, are not semantically complex. Take the idiom 'to keep an eye out for.' While the sentence 'I'm keeping an eye out for you' is perfectly linguistically appropriate, its cleft analogue 'It's an eye that I'm keeping out for you' strikes us as odd. This despite the fact that with non-idiomatic expressions, a cleft sentence is very close in meaning to its non-cleft variant; viz. 'I gave her an umbrella' and 'It's an umbrella that I gave her.' The reason for this is that the meaning of 'to keep an eye out for,' unlike the meaning of 'to give an umbrella to,' is not derived compositionally from the meanings of its parts. Rather, its meaning is directly lexically encoded by the whole expression. This means that, on the level of semantics, 'to keep an eye out for' is a single unit that can't be broken up by, say, cleft constructions. In contrast, counterfactuals about relative computability interact with other sentence constructions just like ordinary counterfactuals do. For example, not only is (valid > arith) false, but the following where we add a negation is true:

> (valid > $\overline{\text{arith}}$) (Even) if the validity problem were algorithmically decidable, arithmetical truth would (still) not be algorithmically decidable.[20]

We will see more examples of how these counterfactuals interact with quantifiers and conjunction shortly. From this, it emerges that the compositional behavior of counterfactuals about relative computability is just like that of ordinary counterfactuals, so that they can't be merely idiomatic ways of speaking.

*Glosses.* A related option would be to treat the counterfactuals used by relative computability theorists as imperfect glosses or paraphrases of claims about reducibility. It may be thought that the case is analogous to the case of causation.[21] When asked to explain what we mean by 'causation,' we make free use of counterfactual locutions. But, so the proposal goes, the failure of the program of analyzing causation in terms of counterfactuals should teach us that we shouldn't take counterfactual locutions as they appear in writings on relative computability too seriously. So, on this proposal, counterfactuals don't characterize or define the reducibility relation, they merely illuminate it.

There are two problems with this analogy with causation. First, the problem with counterfactual analyses of causation is that they notoriously either over- or undergenerate cases of genuine causation. Things are different in the case of relative computability. If we bracket the violations of the vacuity thesis—which it is fair to bracket, since the status of the vacuity thesis is the very thing that's at issue—counterfactual glosses on the notion of reducibility seem to get things *exactly* right.

Secondly, we seem to have an understanding of causation that's independent of our understanding of counterfactuals. In fact, several authors have recently argued that we should give a semantics for counterfactuals in terms of causal models, the latter of which treat causation as a primitive notion.[22] In contrast, it's implausible that the notion of reducibility that's at the core of relative computability theory is primitive. We simply don't have a pretheoretical notion of reducibility that's not understood by way of some auxiliary notions. My present claim is that reducibility is understood in terms of counterfactuals, and that some of these counterfactuals are counterpossibles. It would be entirely mysterious how such an understanding could be achieved if the vacuity thesis were correct. Of course, whether my claim about how we understand reducibility is true will in part depend on whether there are ways of understanding the reducibility relation that don't involve counterpossibles. I will discuss some potential definitions presently.

*Conceptual possibility.* One tempting response to counterpossibles that appear to be non-trivial is to interpret them as talking about what's *conceptually* possible. The notion of conceptual possibility is a famously fraught one, since it is tied to the notions of apriority and analyticity. $\phi$ is sometimes said to be epistemically possible iff $\ulcorner\neg\phi\urcorner$ isn't knowable a priori, and sometimes it's said that $\phi$ is epistemically possible iff $\ulcorner\neg\phi\urcorner$ isn't true in virtue of meaning. We needn't be concerned with the details here. Let's just grant that there is a notion of conceptual possibility according to which it's conceptually possible that water is an element and that Ms. Marvel, the heroine of the eponymous comic book series, isn't Kamala Khan. A conceptually possible world can then be defined as a maximal consistent set of sentences that includes all a priori knowable or analytic truths.[23] Conceptually possible worlds may be used to give a model theory for counterfactuals such as the following:

> (water) If water had been an element, then water splitting would have been impossible.

> (marvel) If Ms. Marvel hadn't been Kamala Khan, we would have seen them together at some point or another.

Since the sentences 'Water isn't an element' and 'Ms. Marvel is Kamala Khan,' though true, are neither a priori knowable nor analytic, there will be conceptually possible worlds where the antecedents of (water) and (marvel) are true. Note that it's crucial for this general strategy to be promising that the building blocks out of which we construct the worlds are sentences, or perhaps Fregean senses, and not something more worldly such as Russellian propositions. The Russellian proposition corresponding to 'Ms. Marvel is Kamala Khan' is the same as the Russellian proposition corresponding to the logical truth 'Ms. Marvel is Ms. Marvel,' and so there isn't any consistent set of Russellian propositions that contains the Russellian proposition corresponding to 'Ms. Marvel is Kamala Khan.'[24]

Given the promise of conceptually possible worlds constructed out of sentences in giving a model theory for (water) and (marvel), it's tempting to also use them to give a model theory for counterfactuals about relative computability. After all, it's plausible that it's conceptually possible that the validity problem is algorithmically decidable.[25] However, conceptual possibility notoriously run into difficulties when it comes to quantifying-in.[26] Indeed, it is commonly assumed that it is illegitimate to quantify into sentential contexts that involve conceptual possibility. But now note that there are certain results about relative computability that require quantifying-in when we express them using counterfactuals. Take Gerald Sacks' (1964) Density Theorem. It states that the Turing reducibility relation is dense.[27] Where $A \leq_T B$ says that $A$ is Turing reducible to $B$ and $A <_T B$ says that $A \leq_T B$ and $B \not\leq_T A$, this theorem can be expressed as follows: for any two sets $A$, $B$, if $A <_T B$, there is a set $C$ such that $A <_T C <_T B$. Using the Post-Turing thesis, we can express this theorem as follows:

> (sacks) For any $A$, $B$, if it's the case that $A$ would be computable if $B$ were computable but not vice versa, then there's some $C$ such that: $A$ would be computable if $C$ were computable but not vice versa and $C$ would be computable if $B$ were computable but not vice versa.

So we see that quantifying-in is very natural for counterfactuals about relative computability. This sets these counterfactuals apart from the kinds of examples commonly discussed in the literature on counterpossibles.[28]

Note that the present claim isn't that the fact that sentences such as (sacks) involve quantification into counterfactuals prohibits the use of *any* world-like entities in their analysis. In fact, the model theory I will present later also involves world-like entities. The present claim is just that the presence, and indeed indispensability, of quantifying-in in some counterfactuals about relative computability calls for more elaborate resources than just conceptually possible worlds qua maximal consistent sets of sentences.

*Semantic ascent.* Perhaps counterfactuals about relative computability are best understood as making meta-linguistic remarks about the predicate 'algorithmically

decidable': (valid > halt) says that if the extension of 'algorithmically decidable' had included the validity problem, then it would also have included the halting problem. In discussing a proposal like this, Berit Brogaard and Joe Salerno (2013) assert that they 'highly doubt that there is an elegant and convincing pragmatic story to be told' about why we would ascend semantically in such a way (p. 645). Contrary to this, I submit that we can tell at least a partial story using Stalnaker's (1978) apparatus of diagonalization. Without going into too many details, this apparatus could be extended quite straightforwardly to predict that counterfactuals with impossible antecedents receive a non-standard reading on which they make meta-linguistic remarks such as the above.[29]

Nevertheless, this story would remain incomplete. Suppose for simplicity that we give a simple Stalnakerian semantics for the reinterpreted counterfactual: the closest world where the extension of 'algorithmically decidable' includes the validity problem is such that at that world, the extension also includes the halting problem. We may ask why this would be so. Surely, the extension of 'algorithmically decidable' could have differed in all sorts of ways. For example, the minimal way of changing the extension so as to include the validity problem would be to *just* add the validity problem and nothing else. Surely, it isn't a brute fact about the predicate 'algorithmically decidable' that this minimal change isn't what happens at the closest world. The reason as to why this minimal change is ruled out must lie in the fact that the halting problem is reducible to the validity problem. But now we're taking the notion of reducibility as more basic than the counterfactuals in terms of which we had originally defined that notion. So now it looks like the best we can do to explain why the closest world where the extension of 'algorithmically decidable' includes the validity problem is such that the extension also includes the halting problem is by appealing to the truth of (valid > halt). This suggests that we have a better grip on the literal interpretation of (valid > halt) than on its meta-linguistic reinterpretation.

*The* reductio *analogy.* Maybe we can understand counterfactuals about relative computability along the lines of counterfactuals found in informal *reductio* proofs.[30] Consider Euclid's proof that there are infinitely many primes. We start by supposing that there are exactly $n$ many primes. Let $p_1, \ldots, p_n$ be them. It follows that there will be a prime $p$ that divides $p_1 \times \ldots \times p_n + 1$. The crucial next step in the proof can then be put in counterfactual terms:

> (euclid) If $p$ were one of $p_1, \ldots, p_n$, then $p$ would divide $(p_1 \times \cdots \times p_n + 1) - p_1 \times \cdots \times p_n$.

Since nothing divides $(p_1 \times \cdots \times p_n + 1) - p_1 \times \cdots \times p_n = 1$, we conclude by *modus tollens* that $p$ isn't one of $p_1, \ldots, p_n$, and so that $p_1, \ldots, p_n$ aren't all of the primes after all. Now, there is some debate over whether counterfactuals such as (euclid) pose a serious challenge to the standard approach to counterfactuals.[31] Suppose they don't. And suppose that counterfactuals such as (euclid) are best understood either as material conditionals or as strict conditionals. This may be particularly plausible in cases where the material conditional is a logical truth,

for in that case a normal modal logic proves the corresponding strict conditional, and both Stalnaker's and Lewis' counterfactual logics then prove the corresponding counterfactual. In any case, whatever the details of the story may be that we tell about (euclid), the present proposal on behalf of the orthodoxy suggests that we treat counterfactuals about relative computability along the same lines. (valid > halt) and (valid > arith), the proposal goes, are merely disguised material or strict conditionals.

The problem with this proposal is that counterfactuals about relative computability don't behave like material or strict conditionals. The reason why Stalnaker and Lewis developed their model theory for counterfactuals is that natural language counterfactuals fail to conform to antecedent strengthening, which is valid for material and strict conditionals. Focusing on the case of strict conditional, this principle reads:

$$\frac{\Box(\phi \rightarrow \psi)}{\Box((\phi \wedge \chi) \rightarrow \psi)}$$

This rule seems adequate for (euclid). No matter what else we put in its antecedent to strengthen it, the resulting sentence still seems true, though perhaps misleading.[32] However, there are counterexamples to antecedent strengthening in the case of counterfactuals about relative computability. Consider:

> (valid > $\overline{\text{arith}}$) (Even) if the validity problem were algorithmically decidable, arithmetical truth would (still) not be algorithmically decidable.
> (valid & arith > $\overline{\text{arith}}$) (Even) if the validity problem and arithmetical truth were algorithmically decidable, arithmetical truth would (still) not be algorithmically decidable.

On the strict conditional interpretation, the inference from (valid > $\overline{\text{arith}}$) to (valid & arith > $\overline{\text{arith}}$) is an instance of antecedent strengthening. But (valid > $\overline{\text{arith}}$) is true and (valid & arith > $\overline{\text{arith}}$) is false.

In response, it may be suggested that the negation in (valid > $\overline{\text{arith}}$) and (valid & arith > $\overline{\text{arith}}$) is a wide-scope negation so that (valid > $\overline{\text{arith}}$) and (valid & arith > $\overline{\text{arith}}$) become '$\neg\Box(V \rightarrow A)$' and '$\neg\Box((V \wedge A) \rightarrow A)$' respectively. Perhaps some story can be told according to which the added 'even' and 'still,' which make (valid > $\overline{\text{arith}}$) and (valid & arith > $\overline{\text{arith}}$) sound more natural, force such a wide-scope interpretation.[33] On this regimentation, the inference from (valid > $\overline{\text{arith}}$) to (valid & arith > $\overline{\text{arith}}$) isn't an instance of antecedent strengthening anymore.

However, this response won't work in full generality. For consider:

> (valid > halt & $\overline{\text{arith}}$) If the validity problem were algorithmically decidable, then the halting problem would be algorithmically decidable but arithmetical truth would (still) not be algorithmically decidable.
>
> (valid & arith > halt & $\overline{\text{arith}}$) If the validity problem and arithmetical truth were algorithmically decidable, then the halting problem would be

> algorithmically decidable but arithmetical truth would (still) not be algorithmically decidable.

As before, (valid $>$ halt & $\overline{\text{arith}}$) is true but (valid & arith $>$ halt & $\overline{\text{arith}}$) is false. But here, there is no temptation whatsoever to treat the negation embedded within the consequent as taking wide scope over the whole counterfactual.

What's more, there are even counterexamples to the claim that a negation that appears unembedded in the consequent of a counterfactual should always be read as taking wide scope. It follows from Corollary 1 in §2.2 of Kleene and Post (1954) that there are sets of natural numbers $A$ and $B$ neither of which is reducible simpliciter to the other. This means that we should be inclined to reject the following:

> ($A \vee B > \overline{B}$) If $A$ or $B$ were algorithmically decidable, then $B$ wouldn't be algorithmically decidable.

But now if it were mandatory to read the negation in ($A \vee B > \overline{B}$) as taking wide scope, we should expect to accept the following:

> ($A \vee B > B$) If $A$ or $B$ were algorithmically decidable, then $B$ would be algorithmically decidable.

In fact, however, we should reject ($A \vee B > B$) for the same reason that leads us to reject ($A \vee B > \overline{B}$).[34]

In short, the claim that counterfactuals about relative computability are material or strict conditionals is untenable.[35]

*The primacy of oracles.* Here's a proposal on behalf of the orthodoxy that exploits the particulars of what these counterfactuals are about. The proposal is that, for example, (valid $>$ halt) is merely shorthand for saying that if we had an oracle for the validity problem, then we could figure out the right answer to any question we may ask about the halting problem.[36] This proposal is inspired by the way we study relative computability, namely by way of oracle Turing machines. What may further motivate this proposal is the thought that there isn't a clear phenomenon, relative computability, that we have a grasp of independently of studying it with oracle Turing machines. Perhaps all we have in relative computability theory is a mathematically rich and thus mathematically interesting structure that doesn't correspond to anything non-mathematical. Don't we all know that mathematicians can become interested in just about any arcane phenomenon as long as it gives rise to a mathematically interesting structure? What's more, understanding counterfactuals such as the above as merely shorthand for saying that we had an oracle for the validity problem would make its antecedent metaphysically possible. For certainly, the proposal continues, though perhaps nomically impossible, oracles by themselves surely aren't metaphysically impossible. Perhaps there could have popped up out of nowhere an oracle that intuits facts about the validity problem. In fact, look back at the quotes from Davis and Enderton (see pages 2 and 3). Davis' counterfactual begins with 'If we could solve $P$...' and Enderton's begins with 'If, hypothetically speaking, we could somehow decide membership in $B$...' Regarding the quote

from Davis, I said that to solve a problem just is to algorithmically decide it. Perhaps I was too quick here. Perhaps Davis has in mind a more general notion of solving, and Enderton has in mind a more general notion of deciding, one that allows reference to metaphysically possible oracles that pop up out of nowhere. A more sober rendition of the present proposal is the following:

> (info) When relative computability theorists assert (valid > halt), what they're *really* doing is assert that there is an algorithm that would allow us to decide which natural numbers are members of the halting problem if we were given complete information about which natural numbers are members of the validity problem.[37]

Just like oracles are metaphysically possible, it's metaphysically possible to be given complete information about which natural numbers are members of the validity problem.

The claim that there isn't any phenomenon to be studied that we understand independently of the notion of an oracle Turing machine runs directly counter to how Rogers develops subject in his book. In chapter 8, Rogers describes a relation of 'reducibility' (the scare quotes are Rogers') among sets that is similar to Turing reducibility, called truth-table reducibility, but which is not defined in terms of oracle Turing machines. After describing truth-table reducibility, Rogers argues for the need for the stronger relation of Turing reducibility in chapter 9, which of course is defined in terms of oracle Turing machines. His argument goes as follows. He produces two sets, the first of which he argues is reducible to the second. He then shows that the first set isn't *truth-table* reducible to the second, but that it is *Turing* reducible to it. Rogers concludes that using truth-table reducibility to analyze what he explicitly calls the *intuitive* notion of reducibility would be inadequate, for this would leave out certain sets, and that an analysis in terms of Turing reducibility fares better. To arrive at this verdict, Rogers clearly assumes that he and his readers have an understanding of the notion of reducibility that's independent of talk about oracle Turing machines. And the understanding of reducibility that Rogers provides is in terms of counterfactuals. In fact, looking back at his quote reveals that it's more difficult to read Rogers in such a way that he's talking about something metaphysically possible. For Rogers' (syntactically non-standard) counterfactual begins with 'given any method for calculating [the characteristic function of B]. . .' And simply being given information doesn't involve any calculating; after all, calculating the validity problem is metaphysically impossible.

Note that the present claim isn't that Rogers assumes that his use of counterpossibles allows himself and his readers to gain an explicit knowledge of the full extension of the relation of reducibility and that he then holds up this extension against the extension of Turing reducibility. Rather, the claim is that Rogers assumes that his use of counterfactuals allows himself and his readers to have an implicit grasp of the notion of reducibility. It may well be, and in fact it is quite plausible, that to pin down the exact boundaries of the extension of the relation of reducibility, the notion of Turing reducibility, which allows for a precise mathematical analysis,

is indispensable. But to admit this is consistent with claiming that counterpossibles are essential in pinning down the subject matter of relative computability theory.[38]

Regarding the analysis of reducibility in terms of (info), I don't deny that this analysis succeeds, just like I don't deny that the analysis of reducibility simpliciter in terms of Turing reducibility succeeds. However, (info) crucially appeals to the notion of a *relative* algorithm, i.e. an algorithm that is given complete information about a certain set of natural numbers. While working relative computability theorists of course have a grasp of this notion, the fact that Rogers sees the need to introduce the notion of reducibility in terms of counterpossibles that don't appeal to the notion of a relative algorithm suggests that the conceptual building blocks that are required for an understanding of relative computability theory are the notion of a non-relative algorithm on the one hand and counterfactuals on the other. But these building blocks only succeed in facilitating an understanding of relative computability theory if the vacuity thesis is false.

That Rogers assumes that he and his readers come to have an understanding of reducibility by way of his use of counterpossibles may be dismissed if Rogers were a minor figure in relative computability theory and if his readers were few. However, from its initial release in 1967 until at least the release of Robert Soare's (1986) textbook, Rogers' book was *the* main textbook with the help of which a whole generation of mathematicians was raised.

*The primacy of hypercomputers.* Another topic specific proposal suggests that the study of relative computability is the study of metaphysically possible *hypercomputers*. Hypercomputers are hypothesized machines that overcome the finiteness of actual computers in one way or another. Many such machines have been described in the literature.[39] One is a so called *accelerating Turing machine*, also called *Zeus machine*.[40] This is a machine that completes an infinite number of computational steps in a finite amount of time. One way it could do this is by completing a supertask, e.g. by completing the first computational step in one minute, the second step in half a minute, the third step in fifteen seconds, and so on. In other words, the machine completes each computational step after the first one in half the time it took to complete the previous one. After two minutes have passed, the machine will have completed an infinite number of steps. There's some debate about whether accelerating Turing machines and the supertasks that they require are physically possible.[41] But they surely seem to be *metaphysically* possible.[42] Now, accelerating Turing machines could 'decide' the validity problem. That's because that set, though algorithmically undecidable, is *computably enumerable*. This means that the set of predicate logic validities is such that a Turing machine, given an infinite amount of time, could list all of its members. Consequently, an accelerating Turing machine of the sort described above could list all and only the members of that set in two minutes. In order to decide in a finite amount of time whether a sentence of the predicate calculus is logically valid, this machine would then just have to generate the list and determine whether the sentence appears on it or not.

So perhaps talk about relative computability could be cashed out in terms of talk about hypercomputers: a set $A$ is reducible to a set $B$ iff: if the laws of nature

allowed for a hypercomputer that could decide membership in $B$, then the laws would also allow for a hypercomputer that could decide membership in $A$. The only modalities involved here are metaphysical.

However, this proposal makes false predictions. It predicts that there will be true counterfactuals of the form,

> ($A \leq_{hyper} B$) If the laws of nature would allow for a hypercomputer that could decide membership in $B$, then the laws would also allow for a hypercomputer that could decide membership in $A$,

even though the corresponding claim about Turing reducibility,

> ($A \leq_T B$) $A$ is Turing reducible to $B$,

is false. To see this, note that there are some algorithmically undecidable but computably enumerable sets $A$ and $B$ that are such that it's neither the case that $A$ is reducible to $B$ nor vice versa.[43] Since $A$ is computably enumerable, a Zeus machine could 'decide' $A$. But presumably if the laws of nature allowed for there to be a Zeus machine that could 'decide' $A$, then they would also allow for there to be a Zeus machine that could 'decide' $B$, since $B$ is computably enumerable as well. In other words, if we had a hypercomputer to decide membership in $A$, then we could also have a hypercomputer to decide membership in $B$. Consequently, ($A \leq_{hyper} B$) is true, even though ($A \leq_T B$) is false. So the explanation of claims involving the Turing reducibility relation, and in turn of relative computability, in terms of what a Zeus machine could do yields the wrong results. And in fact, according to the theory of supertask computation as developed by Joel David Hamkins (2004) and Philip Welch (2004), Zeus machines are vastly more powerful than many oracle machines.[44]

The failure of these seven proposals suggests that the reducibility relation is indeed to be understood in terms of counterpossibles. This means that the orthodoxy about counterfactuals does indeed rob relative computability theory of its subject. Philosophical humility thus recommends that we reject the orthodoxy. But perhaps we think that philosophical humility has its limits. Perhaps we want to dig in our heels and insist that counterfactuals such as ($\mathsf{valid} > \mathsf{halt}$) and ($\mathsf{valid} > \mathsf{arith}$) are indeed both true. This attitude owes us a story as to why these counterfactuals strike us as *prima facie* non-vacuous. We can take a cue from Williamson's (2015) discussion here.[45]

The following is a version of an argument of Williamson's that purports to put pressure on our inclination to treat ($\mathsf{valid} > \mathsf{arith}$) as false using general principles of the logic of counterfactuals. The argument, which is adapted for our purposes, starts by claiming that ($\mathsf{valid} > \mathsf{arith}$) is equivalent to ($\mathsf{valid} \,\&\, \overline{\mathsf{valid}} > \mathsf{arith}$):

> ($\mathsf{valid} \,\&\, \overline{\mathsf{valid}} > \mathsf{arith}$) If the validity problem were and weren't algorithmically decidable, then arithmetical truth would be algorithmically decidable.

Why should this equivalence hold? It's metaphysically necessary that the validity problem isn't algorithmically decidable. And since $\phi$ is metaphysically equivalent to $\ulcorner \phi \wedge \psi \urcorner$ whenever $\psi$ is metaphysically necessary, 'the validity problem is

algorithmically decidable' is metaphysically equivalent to 'the validity problem is and isn't algorithmically decidable.' In worlds talk, 'the validity problem is algorithmically decidable' is true at all the same metaphysically possible worlds as 'the validity problem is and isn't algorithmically decidable.' Next, it is claimed that counterfactuals allow for substitution of necessary equivalents; i.e. if $\phi$ and $\psi$ are true at all the same metaphysically possible worlds, then $\ulcorner \phi \mathbin{\square\!\!\rightarrow} \chi \urcorner$ and $\ulcorner \psi \mathbin{\square\!\!\rightarrow} \chi \urcorner$ are equivalent. This gives us the desired equivalence between (valid > arith) and (valid & $\overline{\text{valid}}$ > arith). Now, surely (valid & $\overline{\text{valid}}$ > arith), with its logically impossible antecedent, is much less obviously false than (valid > arith). So perhaps we are merely tricked into thinking that (valid > arith) is false because we don't realize that it's equivalent to (valid & $\overline{\text{valid}}$ > arith).

However, a closer look at this argument reveals that it rests on an assumption that we ought to reject for the same reason that we ought to accept counterfactuals about relative computability as non-vacuous. Let's look at how we would derive the supposed equivalence between (valid > arith) and (valid & $\overline{\text{valid}}$ > arith). Stalnaker's (1968, 106) counterfactual logic *C*2 contains the following axioms:

**(a3)** $\square(\phi \rightarrow \psi) \rightarrow (\phi \mathbin{\square\!\!\rightarrow} \psi)$

**(a7)** $\big((\phi \mathbin{\square\!\!\rightarrow} \psi) \wedge (\psi \mathbin{\square\!\!\rightarrow} \phi)\big) \rightarrow \big((\phi \mathbin{\square\!\!\rightarrow} \chi) \leftrightarrow (\psi \mathbin{\square\!\!\rightarrow} \chi)\big)$

Now, since it's metaphysically necessary that the validity problem isn't algorithmically decidable, we have:

$$\square\Big(V \leftrightarrow (V \wedge \neg V)\Big)$$

Using (a3), this gives us:

$$\Big(V \mathbin{\square\!\!\rightarrow} (V \wedge \neg V)\Big)$$

and

$$\Big((V \wedge \neg V) \mathbin{\square\!\!\rightarrow} V\Big)$$

And so (a7) gives us:

$$\Big(V \mathbin{\square\!\!\rightarrow} A\Big) \leftrightarrow \Big((V \wedge \neg V) \mathbin{\square\!\!\rightarrow} A\Big)$$

That (a3) gives us '$V \mathbin{\square\!\!\rightarrow} (V \wedge \neg V)$' is suspicious. For it follows from this that any counterfactual that assumes in its antecedent that the validity problem is algorithmically decidable is vacuous. And whether that's the case is exactly what's at issue. So if '$\square$' is interpreted as metaphysical necessity, then we ought to reject (a3). In counterfactual logic, '$\square$' is usually defined such that $\ulcorner \square\phi \urcorner$ abbreviates $\ulcorner \neg\phi \mathbin{\square\!\!\rightarrow} \phi \urcorner$. That's how Lewis (1973, §1.5) defines it; he calls '$\square$' *outer necessity*. (a3) is valid in the model theory I present in the appendix if that's how we understand '$\square$,' since outer necessity is now broader than metaphysical necessity. But if that's how we understand '$\square$,' then we can't accept '$\square\big(V \leftrightarrow (V \wedge \neg V)\big)$.' The latter is true only where '$\square$' is understood as metaphysical necessity. So the logic of counterfactuals

doesn't force upon us the equivalence of (valid > arith) and (valid & $\overline{\text{valid}}$ > arith). And without this equivalence, it becomes less plausible that we are tricked into thinking that (valid > arith) is false. Note that given this notion of outer necessity, the debate over the vacuity thesis can be rephrased as follows: is outer necessity the same as metaphysical necessity? Stalnaker, Lewis, and Williamson think that it is, whereas I argue that outer necessity is stronger than metaphysical necessity.

A final way of holding on to the orthodoxy is to argue that despite its shortcomings, it's the only game in town, since all alternative approaches such as for example that of Brogaard and Salerno (2013) run into serious trouble. And indeed, perhaps there's a way of amending the orthodoxy by providing an error theory about our judgments about counterpossibles. Williamson (2015, §4), for example, proposes that we use certain heuristics when evaluating counterfactuals that lead us astray in cases of counterpossibles. However, in the next section, I describe a model theory for counterfactuals about relative computability, which I describe in more detail in the appendix, which I hope demonstrates that the orthodoxy isn't the only game in town.

## 5. Patching Up the Orthodoxy

Williamson likens the supposed folly of rejecting the vacuity thesis to the Aristotelian rejection of vacuously true universal generalizations:

> The logic of quantifiers was confused and retarded for centuries by unwillingness to recognize vacuously true universal generalizations; we should not allow the logic of counterfactuals to be similarly confused by unwillingness to recognize vacuously true counterpossibles. (Williamson 2007, 175)

Given the fact that the standard model theory of counterfactuals treats counterfactuals as universal quantifiers over worlds, Williamson's analogy is of course particularly apt. Do we risk entering a kind of logical Dark Age if we accept that counterfactuals such as (valid > halt) and (valid > arith) are non-vacuous? Fortunately, there is no such risk. On the model theory for counterfactuals about relative computability presented in the appendix, these counterfactuals are still universal quantifiers over indices and they still admit of vacuously true instances. In fact, the model theory is of a piece with Lewis' similarity models; it incorporates a version of the vacuity thesis insofar as it treats counterfactuals with outright logical falsehoods in the antecedents as vacuously true.

The basic idea of the model theory is simple. Relative computability theory provides us with an abstract structure called the *Turing degrees*. Informally, we can say that this structure classifies sets of natural numbers into complexity classes. The halting problem and the validity problem belong to the same complexity class, which is why (valid > halt) and its converse are true, but arithmetical truth belongs to class of problems of much higher complexity, which is why (valid > arith) is false. The Turing degrees form a hierarchy that has the form of an infinite tree originating from a single point.[46] This point of origin is the class of least complex sets, i.e. the sets that are in fact computable. For example, the set $\omega$ of all natural numbers belongs to this

class, since we can easily come up with an algorithm for deciding it: for any number $n$, to decide whether $n$ is in $\omega$, compute nothing and output 'yes.' Another way of thinking of this least class is that it represents something like the actual world: everything that's actually algorithmically decidable is represented by this class as algorithmically decidable. This is the class where the Church-Turing holds and so where the laws of computation are as they actually are. So it's tempting to just have the Turing degrees play the role of worlds, where all of the Turing degrees except for the one that stands for the actual one are thought of as non-actual worlds where the laws of computation are different. The further you move up the tree, the more violations of the Church-Turing thesis you get, since more and more sets that aren't actually algorithmically decidable become represented as algorithmically decidable. This tree-like structure gives us everything we need for Lewis' comparative similarity semantics for the counterfactual connective. Unfortunately, this isn't quite right, for reasons explained in the appendix. What we need for our worlds are rather *ideals* on Turing degrees. The ideals still form a tree-like structure on which we can build Lewis' comparative similarity semantics. A simple counterfactual ⌜If $B$ were algorithmically decidable, then $A$ would be algorithmically decidable⌝ is true at a world $w$ (i.e. an ideal on Turing degrees) iff all worlds closest to $w$ that represent $B$ as algorithmically decidable also represent $A$ as algorithmically decidable.[47] We can turn this into a fully general semantics for the counterfactual connective by incorporating the standard semantic clauses for the Boolean connectives and the quantifiers. As long as the semantic clauses for the connectives are classical, $\ulcorner(\phi \wedge \neg\phi)\Box\!\!\rightarrow \psi\urcorner$ comes out vacuously true, for any $\psi$, since there's no world where $\ulcorner\phi \wedge \neg\phi\urcorner$ is true. Again, for more details, see the appendix, and for a complete axiomatization of a propositional fragment of what I call *the conditional logic of Turing reducibility*, see Jenny (MS).

Let's take stock. Not only do we have positive reasons for interpreting counterfactuals about relative computability literally, as seen in the previous section, but we can also see now that nothing stands in the way of extending Lewis' similarity models to give a model theory for these counterfactuals. The resulting theory doesn't have us falling back into a logical Dark Age that Williamson has warned us of. Our job isn't done, however. One big remaining question is how to interpret our model theory. Even though the ideals on Turing degrees in the model theory just sketched act like worlds as far as the model theory is concerned, they are of course a very different kind of object than what we usually think of when we think of worlds, possible or impossible. I take up this issue in the next section.

## 6. Interpreting the Indices

The reason why the ideals on Turing degrees, which are just sets of sets of sets of natural numbers, act like worlds as far as the above model theory is concerned says more about the model theory than about the ideals. As is well known, so-called 'possible worlds' model theory doesn't presuppose any kind of realism about possible worlds. As a piece of mathematics, the model theory doesn't care what the

'worlds' are that we use. These worlds are just indices at which we evaluate sentences. So there's nothing mysterious about the fact that ideals on Turing degrees can act as indices.

However, we may still ask what possible worlds model theory is *for*, and depending on what we think it's for, we may want to ask some more probing questions about how to interpret the role of the ideals on Turing degrees in the above model theory. Of course, it is beyond the scope of this paper to develop a theory of model theory. But I want to make a few remarks about how my proposed model theory fits into two alternative pictures of the role of model theory.

On an *instrumentalist* understanding of possible worlds model theory, possible worlds models are merely a useful tool to study the logic of the object languages in question. There's no doubt that possible worlds model theory has greatly advanced our understanding of modal and counterfactual logic. But an appreciation of the usefulness of model theory is consistent with the rejection of any sort of realism about possible worlds. One form of such instrumentalism is *modalism*.[48] Modalism claims that the modal operators and counterfactual connectives are in some sense more basic than the possible worlds used in their model theory. Kit Fine (1977) explicitly speaks of the *construction* of possible worlds. So the rough idea is that we 'construct' possible worlds using our modal and counterfactual language and then use them to obtain a more precise understanding of that language. This take on possible worlds model theory fits particularly well with the way we use the ideals on Turing degrees in the above model theory. For after all, as described in the appendix, the ideals on Turing degrees are selected from among the mathematical universe to play the role of worlds with the help of the Turing reducibility relation. The Turing reducibility relation in turn corresponds to the relation of reducibility simpliciter, via the Post-Turing thesis. And as we've seen, reducibility simpliciter is best cashed out in counterfactual talk. So on a modalist-instrumentalist understanding of possible worlds model theory, there is no puzzle about the role of the ideals on Turing degrees in our model theory.

There is also a more inflationary understanding of possible worlds model theory, the *representational* understanding. The idea here is that there is a privileged possible worlds model, the one that corresponds to the semantics of our language, and that model captures the truth conditions of our sentences.[49] Such a representational understanding of course presupposes a kind of realism about possible worlds. But that realism needn't be as strong as Lewis' (1986); a weaker realism, such as perhaps Stalnaker's (2003; 2012), suffices.[50] Given such realism, the question how the ideals on Turing degrees *qua* indices relate to possible worlds becomes pressing. Whatever possible worlds are, they surely aren't sets of sets of sets of natural numbers. So if we want to give genuine truth conditions for counterfactuals about relative computability, an appeal to ideals on Turing degrees is bound to be unilluminating. However, a representational understanding of our model theory may be available. Suppose there are worlds, possible or otherwise, where the laws of computability are different from what they actually are. And suppose that for any set that appears somewhere in the structure of the ideals on Turing degrees, there's such a world where that set is algorithmically decidable. Then we can define a partition

on the set of all of these worlds such that two worlds are in the same cell iff they agree on the laws of computability. We will then be able to define a model that's isomorphic to the model I present in the appendix where the indices are the cells of the partition. What's more, the Post-Turing will guarantee that the resulting truth conditions for sentences such as (valid > halt) and (valid > arith) will be adequate. And if we want to provide an intended model for a language in which we can talk about more than just algorithmic decidability, we can take this new model and extend the comparative similarity relation to the members of the cells of the partition. This will allow us to assign truth-conditions to counterfactuals whose component sentences talk both about algorithmic decidability as well as about all things other than algorithmic decidability. Of course, this may lead us to assign truth conditions to counterfactuals that involve odd, gerrymandered pairings of sentences about algorithmic decidability and sentences having nothing whatsoever to do with algorithmic decidability. But we can of course have counterfactuals with similarly odd pairings even in the absence of an ability to talk about algorithmic decidability. Such is the nature of compositionality. Perhaps some such pairings will lead us to adopt, say, a model theory that allows for truth-value gaps so that we aren't required to count every counterpossible as either true or false. But there's no reason for thinking that the introduction of an ability to talk about algorithmic decidability will put any pressure on us to go in for such maneuvers that wasn't already there before.

Of course, some will doubt the intelligibility of metaphysically impossible worlds where the laws of computability are different, given the metaphysical necessity of the Church-Turing thesis. Echoing Bertrand Russell's (1905) and W. V. Quine's (1948) criticisms of Meinongian ontology, Lewis (1986, 7 n. 3) and Stalnaker (1996) are suspicious of *logically* impossible worlds where contradictions hold. They argue as follows: suppose that there's an impossible world $w$ at which '$p$' and '$\neg p$' are true. Then given that '$\neg p$' is true at $w$, it's not the case that '$p$' is true at $w$. So it both is and isn't the case that '$p$' is true at $w$. Contradiction. So $w$ can't exist. Whatever the force of this objection may be, it clearly doesn't apply to the present use of metaphysically impossible worlds. For none of the worlds required by our model theory are logically impossible.[51] And clearly, a version of the Stalnaker-Lewis argument against our impossible worlds won't go through. Essentially, we are saying that there are worlds where the Church-Turing thesis fails. To get a contradiction from this, we would need the assumption that the Church-Turing thesis holds in every world. But all I've argued is that the Church-Turing thesis holds in every *metaphysically possible* world. More generally, if we're representationalists about worlds model theory, then our metaphysically impossible worlds earn their keep for much the same reason that metaphysically possible worlds earned their keep: as we saw, they allow us to develop truth conditions for a certain class of counterfactuals.

We thus see that no matter whether we're instrumentalists or representationalists about our model theory, there's no serious worry about its use of indices that represent the laws of computation as different from what they actually are.

## 7. Conclusion

The case for the vacuity of counterfactuals about relative computability looks feeble. We've seen that the reducibility relation, which is the subject of study of relative computability theory, is to be understood in terms of counterfactuals. These counterfactuals have metaphysically impossible antecedents, and so the vacuity thesis threatens to undermine a whole mathematical discipline. Philosophical humility recommends that we revise our theory of counterfactuals before we propose to put our colleagues in mathematics out of a job.

Some questions still remain, however. First, the representational understanding of worlds model theory gives rise to general questions about the metaphysics of worlds, and about whether metaphysically possible worlds are the same kind of thing as metaphysically impossible worlds. These questions are beyond the scope of the present paper.

Another question concerns the status of the outer necessity operator that I briefly discussed at the end of section 4. Is there a theoretically important modality corresponding to this operator that's of the same kind as metaphysical necessity, though more strict? Accepting an ideology of outer necessity would arguably be the most conservative way of amending the orthodoxy, since it would allow us to hold on to a version of the vacuity thesis. In fact, if we accept this ideology, then counterfactuals about relative computability turn out not to be counterpossibles at all, at least not as far as outer possibility is concerned. Whether the ideology of outer possibility is worth accepting for this and other reasons will have to be judged against the same kind of criteria that are used to answer questions about ideological commitment in general.

Finally, one may wonder how the theory I've developed extends to counterpossibles that aren't about relative computability, such as perhaps (water) and (marvel) mentioned on pages 10 and 11. I submit that my discussion gives us reason to take seriously the suggestion, made of course by many in the literature, that there are other non-vacuous counterpossibles. But I also hope that my discussion has shown that careful investigation is required to establish that a given counterpossible is indeed non-vacuous. In particular, since many of the purported counterexamples to the vacuity thesis mentioned in the literature such as those involving claims about what would have happened if the laws of metaphysics had failed are about philosophical topics, an appeal to philosophical humility such as the one I invoke above may not always be available. What we have here is a classic case where one philosophical domain, i.e. metaphysics, is in tension with another, i.e. philosophical semantics. To move beyond the gridlock in the debate over counterpossibles, we need to look for uses of counterpossibles outside of philosophy. I therefore suggest that we seek to find established scientific disciplines other than relative computability theory where counterpossibles play an ineliminable role. It is my hope that the present study has taken a first step towards such a case-by-case study of counterpossibles. Once we have a clearer picture of the areas where non-vacuous counterpossibles are indispensable and once we have model theories for these various classes of counterpossibles, we may then investigate to what extent we can

integrate these model theories to come up with a unified and fully general theory of non-vacuous counterpossibles.

## Appendix: Model Theory

In this appendix, I describe a model for a quantified language of relative computability with a designated predicate '$D$' for algorithmic decidability.[52]

The Turing degree of some set $A$ is $\deg(A) = \{B : A \leq_T B \text{ and } B \leq_T A\}$. We can define an ordering $\leq$ on the Turing degrees $\boldsymbol{D}$ so that for $\mathbf{a}$, $\mathbf{b}$ Turing degrees, $\boldsymbol{a} \leq \boldsymbol{b}$ iff there's some $A \in \boldsymbol{a}$ and some $B \in \boldsymbol{b}$ such that $A \leq_T B$. Informally, the Turing degree of $A$ is its complexity class. I mentioned that it's tempting to think of Turing degrees as worlds, where a degree-world would represent a set as decidable iff it contains that set. However, this would mean, for example, that there is no world where the decidable sets are all *and only* the arithmetically definable sets. This follows from Corollary 1 of §4.4 in Kleene and Post (1954) that there's no degree that contains all *and only* the arithmetical sets, since $\mathbf{0}^{(\omega)}$ isn't a minimal upper bound to the arithmetical degrees $\mathbf{0}, \mathbf{0}', \mathbf{0}'', \ldots$.[53] We can avoid this undesirable result if we use *ideals* on Turing degrees instead. For any $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{D}$ and for $\mathbf{0}$ the degree of the algorithmically decidable sets, an ideal $\boldsymbol{i}$ on the Turing degrees is a non-empty set of Turing degrees such that if $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{i}$, then their join $\boldsymbol{a} \oplus \boldsymbol{b}$ is in $\boldsymbol{i}$ as well; and if $\boldsymbol{a} \in \boldsymbol{i}$ and $\boldsymbol{b} \leq \boldsymbol{a}$, then $\boldsymbol{b} \in \boldsymbol{i}$. Since the join of two arithmetically definable Turing degrees is arithmetically definable and since anything reducible to an arithmetical set is arithmetical, the arithmetical sets form an ideal.

The starting point for our model theory are the frames for Lewis' (1971; 1973) comparative similarity models, which consist of a set of indices (worlds) $\mathfrak{W}$ and a ternary relation $\mathfrak{R}$ on $\mathfrak{W}$ such that for each $w \in \mathfrak{W}$, $\mathfrak{R}_w$ is a total binary preordering on $\mathfrak{W}$. $v\mathfrak{R}_w u$ is informally understood as saying that world $v$ is at least as similar to $w$ as $u$ is to $w$.[54]

The structure of the Turing degrees $\langle \boldsymbol{D}, \leq \rangle$ is very similar to such frames. It is easily seen that $\leq$ partially orders $\boldsymbol{D}$. What is more difficult to see is that $\leq$ isn't total; there are $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\boldsymbol{D}$ such that $\boldsymbol{a} \not\leq \boldsymbol{b}$ and $\boldsymbol{b} \not\leq \boldsymbol{a}$. This is Corollary 1 in §2.2 of Kleene and Post (1954), which we've already encountered. But we already saw that we can't use $\boldsymbol{D}$ to serve as the set of worlds. Rather, we need to use the set $\boldsymbol{I}$ of ideals on Turing degrees. This set already comes partially ordered by the subset relation. But still, a difference between $\langle \boldsymbol{I}, \subseteq \rangle$ (besides the fact that $\subseteq$ isn't total, due to the non-totality of $\leq$) is that $\subseteq$ is a binary relation whereas Lewis' $\mathfrak{R}$ is ternary. This turns out not to be a problem, however.

For $\langle \boldsymbol{I}, \subseteq \rangle$ a frame, our model is the tuple $\mathfrak{M} = \langle \wp(\omega), \boldsymbol{I}, \subseteq, \mathfrak{I} \rangle$, where $\wp(\omega)$ is the power set of the set of natural numbers and $\mathfrak{I}$ takes '$D$' to functions from members of $\boldsymbol{I}$ to subsets of $\wp(\omega)$ such that for $w \in \boldsymbol{I}$ and $x \in \wp(\omega)$, $x \in \boldsymbol{I}(\text{`}D\text{'})(w)$ iff for some $y \in \bigcup w$, $x \leq_T y$. For $g$ a function that assigns members of $\wp(\omega)$ to the variables of the language, we then have that '$Dx$' is satisfied at a world $w$ iff $g(x) \in \mathfrak{I}(\text{`}D\text{'})(w)$.[55] The counterfactual connective '$\boxrightarrow$' is defined as follows (where $\mathfrak{W}_w = \{v \in \mathfrak{W} : w \subseteq v\}$ and $[\![\phi]\!]^g_{\mathfrak{M}}$ is shorthand for $\{w \in \mathfrak{W} : [\![\phi]\!]^g_{\mathfrak{M},w} = 1\}$): $[\![\ulcorner \phi \boxrightarrow \psi \urcorner]\!]^g_{\mathfrak{M},w} = 1$ iff for all $v \in \mathfrak{W}_w \cap [\![\phi]\!]^g_{\mathfrak{M}}$, there is some $u \in \mathfrak{W}_w \cap [\![\phi]\!]^g_{\mathfrak{M}}$ such

that $u \subseteq v$ and such that for any $\in \mathfrak{W}_w$ such that $t \subseteq u$, $[\![\ulcorner \phi \rightarrow \psi \urcorner]\!]^g_{\mathfrak{M},t} = 1$. Note that this clause for '$\Box\!\!\rightarrow$,' which is adapted from Burgess' (1981), differs from Lewis' clause in that it contains an additional initial universal quantifier. This is required because our partial order isn't total, whereas Lewis' comparative similarity relations are. Note also that our binary partial order can be turned into a ternary comparative similarity relation in a canonical way: we define the ternary relation $\subseteq^*$ such that $\boldsymbol{j} \subseteq^*_{\boldsymbol{i}} \boldsymbol{k}$ iff $\boldsymbol{i} \subseteq \boldsymbol{j}$ and $\boldsymbol{j} \subseteq \boldsymbol{k}$. This gives us the frame $\langle \boldsymbol{I}, \subseteq^* \rangle$, on which we can build models each of which belongs to (the quantified version of) John Burgess' (1981) model class $M_1$. If we then redefine $\mathfrak{W}_w$ as $\{v \in \mathfrak{W} : v \subseteq_w v\}$ and take over the above clause for '$\Box\!\!\rightarrow$,' we immediately get that the ternary version of our model on $\langle \boldsymbol{I}, \subseteq^* \rangle$ validates all axioms and rules of Burgess' (1981) logic $S_1$. $S_1$ is strictly weaker than Lewis' (1971) favored counterfactual logic $C1$, which we obtain from $S_1$ by adding:[56]

$$\mathbf{D}'. \Big( (\phi \vee \psi) \Box\!\!\rightarrow \neg\phi \Big) \rightarrow \Big( \big( (\phi \vee \chi) \Box\!\!\rightarrow \neg\phi \big) \vee \big( (\psi \vee \chi) \Box\!\!\rightarrow \neg\chi \big) \Big)$$

And of course from $C1$ we can get Stalnaker's (1968) logic $C2$ by adding conditional excluded middle:

$$\mathbf{CEM}. (\phi \Box\!\!\rightarrow \psi) \vee (\phi \Box\!\!\rightarrow \neg\psi)$$

Neither D′ nor CEM are valid in $\mathfrak{M}$, due to the fact that Corollary 1 in §2.2 of Kleene and Post (1954) makes $\subseteq$ non-total. Regarding the quantifiers, since these models have a fixed domain, the Barcan (1946) formula and its converse come out valid.

Of course, the frame $\langle \boldsymbol{I}, \subseteq \rangle$ has certain features that not all frames have on which the models in Burgess' $\mathcal{M}_{0,1}$ are built. In fact, the structure of the ideals on Turing degrees is an upper semi-lattice with a zero-element, and it has many other features that we may wish to capture axiomatically. I provide a complete axiomatization of a propositional fragment of the conditional logic of Turing reducibility as well as a decision procedure in Jenny (MS). For the quantificational case, we may want to enrich our language with a predicate for computable enumerability and with function signs for the complementation, jump, and join operations on sets of natural numbers. Whether the structure of the ideals on the Turing degrees can be completely axiomatized is unknown. What's important for present purposes is that with our model $\mathfrak{M}$ we have what we need to correctly interpret (regimentations of) our counterfactuals (valid > halt) and (valid > arith) (see page 4), quantified counterfactuals such as (sacks) (page 11), as well as many more.

Before we can regiment (valid > halt) and (valid > arith), we should expand our language to include individual constants '$v$,' '$h$,' and '$a$' that $\mathfrak{I}$ assigns to the validity problem, the halting problem, and arithmetical truth respectively. Then (valid > halt) and (valid > arith) become '$Dv \Box\!\!\rightarrow Dh$' and '$Dv \Box\!\!\rightarrow Da$' respectively. Given that the degree of both the validity and the halting problem is the degree $\mathbf{0}'$ and the degree of arithmetical truth is $\mathbf{0}^{(\omega)}$ and given that $\mathbf{0}' \leq \mathbf{0}^{(\omega)}$, '$Dv \Box\!\!\rightarrow Dh$' comes out true at the zero-element 'world' in $\mathfrak{M}$ and '$Dv \Box\!\!\rightarrow Da$' comes out false,

as desired. In fact, as long as we have a model on the frame $\langle \boldsymbol{I}, \subseteq^* \rangle$ that assigns to the atomic sentences of our language the intended set ideals on Turing degrees, our model theory gives exactly the results we want. For example, it is routine to verify that the relevant regimentations of (valid $>$ $\overline{\text{arith}}$), (valid & arith $>$ $\overline{\text{arith}}$), (valid $>$ halt & $\overline{\text{arith}}$), and (valid & arith $>$ halt & $\overline{\text{arith}}$) (see page 13) have all the desired properties in our model. And since the structure of the Turing degrees is dense, the relevant regimentation of (sacks) is true at any world in our model as well:

$$\forall x \forall y \big( (Dy \Box\!\!\rightarrow Dx) \wedge \neg (Dx \Box\!\!\rightarrow Dy) \big) \rightarrow$$

$$\exists z \big( (Dz \Box\!\!\rightarrow Dx) \wedge \neg (Dx \Box\!\!\rightarrow Dz) \wedge (Dy \Box\!\!\rightarrow Dz) \wedge \neg (Dz \Box\!\!\rightarrow Dy) \big)$$

## Notes

[1] See Nolan (1997) and Brogaard and Salerno (2013) for influential papers and Berto (2013, §5.1) for more references.

[2] See again Berto (2013) for an overview of previous proposals.

[3] See Baras (MS) for a discussion of Brogaard and Salerno's (2013) proposal.

[4] For the sake of simplicity, I am straining traditional usage a bit here. Traditionally, the decision problem was so-called because it called for an algorithm for deciding membership in the set containing the logical validities; it wasn't the set itself that was called 'the decision problem.' See Mancosu and Zach (2015).

[5] See Soare (1996, 2009) for historical overviews of and Soare (2016) for an up-to-date introduction to the theory of relative computability. I will appeal to facts proven in Soare (2016) throughout here. Note that Soare (1996) initiated the change in usage from 'recusion theory' to '(relative) computability theory.'

[6] See Chisholm (1946) and Goodman (1947).

[7] See Stalnaker (1968), Stalnaker and Thomason (1970), and Lewis (1971, 1973). See also Todd (1964) and Sprigge (1970) for early statements of ideas similar to Stalnaker's and Lewis'.

[8] As it happens, the validity problem is also reducible to the halting problem; but the reducibility relation isn't in general symmetrical.

[9] See Shapiro (1981) and Rescorla (2007) for discussions of different interpretations of the thesis.

[10] Of course, talk of such storage devices is purely metaphorical; recall that Turing machines are abstract objects instead of concrete computing devices. So strictly speaking, an oracle is the abstract analogue of a concrete storage device.

[11] See Soare (2009, 382) and Cooper (2004, 142).

[12] See Piccinini (2015, §4.3).

[13] I skip over some complications here; see Piccinini (2015, §4) for a more detailed discussion. In particular, I interpret the Church-Turing thesis as what Piccinini calls the *mathematical* Church-Turing thesis; I take this to be historically accurate. Note also that I discuss the issue of hypercomputation in section 4.

[14] McGee (2006, 111), one of the very few discussions of the modal status of algorithmic decidability, concurs.

[15] See Knuth (2016).

[16] These considerations also suggest that Cleland's (1993) thinking about the Church-Turing thesis is even more revisionary than Cleland herself suggests.

[17] See Kment (2014, 25, 220) for a theory that treats all counterfactuals with *logically* impossible antecedents as vacuously false.

[18] Thanks to Alex Byrne for helpful discussion here.

[19] This attitude is related to Lewis' (1991, §2.8) Credo about set theory and Shapiro's (1997, ch. 1) 'philosophy-last' approach to philosophy of mathematics.

[20] Note that $(\mathsf{valid} > \overline{\mathsf{arith}})$ is only *equivalent* to the negation of $(\mathsf{valid} > \mathsf{arith})$ if we assume conditional excluded middle, which, as we'll see later, is not in general valid. Note also that I assume that 'even' and 'still' don't make any truth-conditional contributions to the counterfactuals in which they occur, which is why I have put them in parentheses; see Bennett (2003, §§102–7) for a defense of this assumption.

[21] Thanks to Bradford Skow for suggesting this analogy and to Justin Khoo and an anonymous referee for urging me to give this proposal more serious consideration.

[22] See Briggs (2012) and the references therein.

[23] I assume here that failures of the laws of logic aren't conceptually possible. This is a harmless assumption in the present context since we're not concerned with counterfactuals with explicit violations of the laws of logic in the antecedent. See Brogaard and Salerno (2013) for an account along the lines I'm imagining here that dispenses with this assumption.

[24] I am grateful to an anonymous referee for pressing me to be clearer on this.

[25] This is assuming, perhaps contrary to Smith (2007, §35), Sieg (2008), and Kripke (2013), that the Church-Turing thesis isn't a conceptual truth. If you disagree, then so much the worse for the present proposal on behalf of the orthodoxy.

[26] These difficulties are most famously noted by Quine (1953).

[27] Strictly speaking, Sacks shows that the relation on the Turing *degrees* is dense. That the Turing reducibility relation is dense is an immediate corollary. For ease of exposition, I'll put off discussion of Turing degrees until the next section.

[28] For example, Brogaard and Salerno (2013) don't even tell us how to extend their model theory to a language with quantifiers.

[29] This strategy may also be carried out using Einheuser's (2012) apparatus.

[30] Thanks to Stephen Yablo for pushing me to think harder about this strategy.

[31] Nolan (1997, 537–8) doesn't think so whereas Dutilh Novaes (forthcoming) does. See also Williamson (2015, §3) for discussion.

[32] If you disagree, then so much the worse for the present proposal on behalf of the orthodoxy.

[33] Though see endnote 20.

[34] It might be worried that we are only inclined to reject $(\mathsf{A} \vee \mathsf{B} > \mathsf{B})$ because simplification of disjunctive antecedents is a valid rule of inference for counterfactuals. This rule reads:

$$\frac{(\phi \vee \psi)\boxrightarrow \chi}{(\phi \boxrightarrow \chi) \wedge (\psi \boxrightarrow \chi)}$$

Simplification isn't valid in Stalnaker's and Lewis' logics of counterfactuals, but Fine (1975, 2012), Ellis et al. (1977), and Santorio (MS) have argued that that's defect of these logics. But even accepting simplification doesn't help in the current situation. For, Kleene and Post's result also leads us to reject the following.

$(\mathsf{A} > \mathsf{B})$ If A were algorithmically decidable, then B would be algorithmically decidable.

But then by simplifiifaction, we should also reject $(\mathsf{A} \vee \mathsf{B} > \mathsf{B})$. So we should accept the negation of $(\mathsf{A} \vee \mathsf{B} > \mathsf{B})$. But then on the assumption that a negation in the consequent of a counterfactual takes wide scope, we should accept $(\mathsf{A} \vee \mathsf{B} > \overline{\mathsf{B}})$ as well, contrary to what we just observed.

[35] von Fintel (2001) and Gillies (2007) have recently argued that natural language counterfactuals only *dynamically* fail to validate antecedent strengthening. Whether they are right is subject to ongoing debate; see Moss (2012) for criticism. But even if von Fintel and Gillies turn out to be right, their dynamic semantics is still very different from the static strict conditional treatment that the current proposal argues is adequate for counterfactuals in *reductio* proofs. For example, von Fintel and Gillies need something like a comparative similarity relation to model the evolution of the context, whereas a static strict conditional treatment only needs an accessibility relation for the modal operators. This also means, in turn, that if we wish to model counterfactuals about relative computability in a dynamic framework, we could just borrow the comparative similarity relation of the model theory that I describe in the appendix.

[36] Agustín Rayo suggested this to me in personal communication.

[37] Thanks to an anonymous referee for suggesting this formulation.

[38] Thanks to an anonymous referee for urging me to be clearer on this.

[39] See Davis (2004) and Piccinini (2015, §4.3) for critical discussions and references.

[40] See Boolos et al. (2007, 19).

[41] See Earman (1995, ch. 4), Davis (2004, 197), and Romero (2014) for discussion.

[42] See Shagrir (2004) for an argument that accelerating Turing machines don't fall prey to Thomson's (1954) paradox.

[43] We know that such sets exists due to Friedberg (1957) and Muchnik's (1956) solution to Post's (1944) Problem.

[44] It may be argued, perhaps with Shapiro (2006), that the informal notion of decidability, and in turn the informal notion of relative computability, can be precisified in a number of ways, one of which coincides with the notion of hypercomputers. But this would still leave us with the result that there's *a* notion of relative computability on which counterfactuals about relative computability are non-vacuous yet have metaphysically impossible antecedents. Thanks to Kieran Setiya for discussion here.

[45] See also Williamson (2010, 95–6) for an earlier discussion.

[46] I use 'tree' in an informal sense here. In its technical sense, trees are well-founded, which the Turing degrees aren't, due to the Sacks Density Theorem.

[47] Since for any set, there's a closest world where that set is algorithmically decidable, this way of glossing the semantic clause is apt. But since the structure of the Turing degrees is dense, a fully accurate statement, such as the one given in the appendix, will have to be slightly more complicated in that we can't rely on the limit assumption.

[48] See Fine (1977), Forbes (1989, 1992), and Williamson (2013, §8.4) for modalism about metaphysical possibility. See also Williamson's (2009, 9) related remarks about the role of Lewis' comparative similarity relation in the analysis of counterfactuals, as well as Stalnaker's (1984, ch. 8) related remarks about the role of his selection functions.

[49] I borrow the expression 'representational' from Etchemendy's (1990, ch. 1) closely related notion of a representational semantics. Note that on a supervaluational treatment of vagueness, we would have a *class* of privileged models, not a single one.

[50] See Berto (2013, §3) for an overview of various theories of possible and impossible worlds.

[51] Of course, if we want to allow for non-vacuous counterfactuals with logically inconsistent antecedents, we will have to face this objection head on. See Berto (2013, §6) for an overview of responses to this objection. In any case, accepting non-vacuous counterfactuals with merely metaphysically impossible antecedents doesn't immediately commit us to such stronger failures of the vacuity thesis.

[52] A construction similar to the present one that's based on the enumeration degrees would yield a model for a language with a designated predicate for computable enumerability. See Odifreddi (1992, ch. XIV) for an introduction to enumeration degrees.

[53] This is how Rogers, Jr. (1967, 276) puts the result in Corollary XVI of §13.4.

[54] Of course, by building on Lewis' model theory, we also inherit some of the potential problems of the Stalnaker-Lewis approach to counterfactuals. For example, it doesn't validate simplification of disjunctive antecedents (see endnote 34). If simplification is indeed desirable, the present model theory can be adapted along the lines developed by Fine (2012) or Santorio (MS) to accommodate it.

[55] Note that for reasons of perspicuity, I use the usual letters '$w$,' '$v$,' '$u$,' and '$t$' to denote 'world' variables here, even though I previously used boldface letters as variables for the members of $\boldsymbol{I}$.

[56] See also Pollock (1976, 43) for a related logic *SS*, which can be turned into *C*1 by adding:

$$((\phi \mathrel{\Box\!\!\rightarrow} \psi) \wedge \neg(\phi \mathrel{\Box\!\!\rightarrow} \neg\chi)) \rightarrow ((\phi \wedge \chi) \mathrel{\Box\!\!\rightarrow} \psi)$$

## References

Baras, D. (MS). 'How close are impossible worlds?'. Unpublished manuscript.

Barcan, R. C. (1946). 'A functional calculus of first order based on strict implication'. *Journal of Symbolic Logic*, 11(1): 1–16.

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

Berto, F. (2013). 'Impossible worlds'. In Zalta, E. N., editor, *Stanford Encyclopedia of Philosophy*. Center for the Study of Language and Information, Stanford University, Stanford, summer 2015 edition.

Boolos, G. S., Burgess, J. P., and Jeffrey, R. C. (2007). *Computability and Logic*. Cambridge University Press, 5th edition.

Briggs, R. (2012). 'Interventionist counterfactuals'. *Philosophical Studies*, 160(1): 139–166.

Brogaard, B. and Salerno, J. (2013). 'Remarks on counterpossibles'. *Synthese*, 190(4): 639–660.

Burgess, J. P. (1981). 'Quick completeness proofs for some logics of conditionals'. *Notre Dame Journal of Formal Logic*, 22(1): 76–84.

Chisholm, R. M. (1946). 'The contrary-to-fact conditional'. *Mind*, 55(220): 289–307.

Church, A. (1936a). 'A note on the *Entscheidungsproblem*'. *Journal of Symbolic Logic*, 1(1): 40–41.

———. (1936b). 'An unsolvable problem of elementary number theory'. *American Journal of Mathematics*, 58(2): 345–363.

Cleland, C. E. (1993). 'Is the Church-Turing thesis true?'. *Minds and Machines*, 3(3): 283–312.

Cooper, S. B. (2004). *Computability Theory*. Chapman and Hall/CRC.

Davis, M. (1958). *Computability and Unsolvability*. McGraw-Hill.

———. (2004). 'The myth of hypercomputation'. In Teuscher, C. (ed.), *Alan Turing: Life and Legacy of a Great Thinker*, Berlin, pages 195–211.

Dutilh Novaes, C. (forthcoming). '*Reductio Ad Absurdum* from a dialogical perspective'. *Philosophical Studies*.

Earman, J. (1995). *Bangs, Crunches, Whimpers, and Shrieks: Singularities and Acausality in Relativistic Spacetimes*. Oxford University Press.

Einheuser, I. (2012). 'Counterconventional conditionals'. *Philosophical Studies*, 127(3): 459–482.

Ellis, B., Jackson, F., and Pargetter, R. (1977). 'An objection to possible-world semantics for counterfactual logics'. *Journal of Philosophical Logic*, 6(1): 355–357.

Enderton, H. B. (2011). *Computability Theory: An Introduction to Recursion Theory*. Elsevier Science.

Etchemendy, J. (1990). *The Concept of Logical Consequence*. Harvard University Press.

Fine, K. (1975). 'Critical notice of David Lewis' *Counterfactuals*'. *Mind*, 84(1): 451–458.

———. (1977). 'Prior on the construction of possible worlds and instants'. In *Modality and Tense: Philosophical Papers*, Oxford, pages 133–175.

———. (2012). 'Counterfactuals without possible worlds'. *Journal of Philosophy*, 109(3): 221–246.

von Fintel, K. (2001). 'Counterfactuals in a dynamic context'. In Kenstowicz, M.J. (ed.), *Ken Hale. A Life in Language*, MIT Press, pages 123–152.

Forbes, G. (1989). *Languages of Possibility: An Essay in Philosophical Logic*. Blackwell.

———. (1992). 'Melia on modalism'. *Philosophical Studies*, 68(1): 57–63.

Friedberg, R. M. (1957). 'Two recursively enumerable sets of incomparable degrees of unsolvability (solution of Post's problem, 1944)'. *Proceedings of the National Academy of Sciences of the United States of America*, 43(2): 236.

Gillies, A. S. (2007). 'Counterfactual scorekeeping'. *Linguistics and Philosophy*, 30(3): 329–360.

Goodman, N. (1947). 'The problem of counterfactual conditionals'. *Journal of Philosophy*, 44(5): 113–128.

Hamkins, J. D. (2004). 'Supertask computation'. In Löwe, B., Piwinger, B., and Räsch, T. (eds.), *Classical and New Paradigms of Computation and Their Complexity Hierarchies* (Papers of the Conference "Foundations of the Formal Sciences III"), Kluwer Academic Publishers, pages 141–158.

Jenny, M. (MS). 'The conditional logic of Turing reducibility'. Unpublished manuscript.

Kleene, S. C. and Post, E. L. (1954). 'The upper semi-lattice of degrees of recursive unsolvability'. *Annals of Mathematics*, 59(3): 379–407.

Kment, B. (2014). *Modality and Explanatory Reasoning*. Oxford University Press.

Knuth, D. E. (1966). 'Algorithm and program; information and data'. *Communications of the ACM* 9(9): 654.

Kripke, S. A. (2013). 'The Church-Turing "thesis" as a special corollary of Gödel's completeness theorem'. In Copeland, B.J., Posy, C.J., and Shagrir, O. (eds.), *Computability: Turing, Gödel, Church, and Beyond*, MIT Press, pages 77–104.

Lewis, D. K. (1971). 'Completeness and decidability of three logics of counterfactual conditionals'. *Theoria*, 37(1): 74–85.
———. (1973). *Counterfactuals*. Harvard University Press.
———. (1986). *On the Plurality of Worlds*. Blackwell.
———. (1991). *Parts of Classes*. Blackwell.
Mancosu, P. and Zach, R. (2015). 'Heinrich Behmann's 1921 lecture on the decision problem and the algebra of logic'. *Bulletin of Symbolic Logic*, 21(2): 164–187.
McGee, V. (2006). 'There are many things'. In Thomson, J.J. and Byrne, A. (eds.), *Content and Modality: Themes from the Philosophy of Robert Stalnaker*, Oxford University Press, pages 93–122.
Moss, S. (2012). 'On the pragmatics of counterfactuals'. *Noûs*, 46(3): 561–586.
Muchnik, A. A. (1956). 'On the unsolvability of the problem of reducibility in the theory of algorithms (Russian)'. *Doklady Akademii Nauk SSSR*, 108(2): 194–197.
Nolan, D. (1997). 'Impossible worlds: A modest approach'. *Notre Dame Journal of Formal Logic*, 38(4): 535–572.
Odifreddi, P. (1992). *Classical Recursion Theory*, volume 2. Elsevier.
Piccinini, G. (2015). 'Computation in physical systems'. In Zalta, E. N., editor, *Stanford Encyclopedia of Philosophy*. Center for the Study of Language and Information, Stanford University, Stanford, summer 2015 edition.
Pollock, J. L. (1976). *Subjunctive Reasoning*. Reidel.
Post, E. L. (1944). 'Recursively enumerable sets of positive integers and their decision problems'. *Bulletin of the American Mathematical Society*, 50(5): 284–316.
Quine, W. V. (1948). 'On what there is'. *Review of Metaphysics*, 2(5): 21–36.
———. (1953). 'Three grades of modal involvement'. In *The Ways of Paradox and Other Essays* (1966), Random House, pages 156–174.
Rescorla, M. (2007). 'Church's thesis and the conceptual analysis of computability'. *Notre Dame Journal of Formal Logic*, 48(2): 253–280.
Rogers, Jr., H. (1967). *Theory of Recursive Functions and Effective Computability*. McGraw-Hill.
Romero, G. E. (2014). 'The collapse of supertasks'. *Foundations of Science*, 19(2): 209–216.
Russell, B. (1905). 'On denoting'. *Mind*, 14(56): 479–493.
Sacks, G. E. (1964). 'The recursively enumerable degrees are dense'. *Annals of Mathematics*, 80(2): 300–312.
Santorio, P. (MS). 'Alternatives and truthmakers in conditional semantics'. Unpublished manuscript, available at http://paolosantorio.net/ac.draft10.pdf, accessed October 21st, 2015.
Shagrir, O. (2004). 'Super-tasks, accelerating turing machines and uncomputability'. *Theoretical Computer Science*, 317(1–3): 105–114.
Shapiro, S. (1981). 'Understanding Church's thesis'. *Journal of Philosophical Logic*, 10(3): 353–365.
———. (1997). *Philosophy of Mathematics: Structure and Ontology*. Oxford University Press.
———. (2006). 'Computability, proof, and open-texture'. In Olszewski, A., Woleński, J., and Janusz, R. (eds.), *Church's Thesis After 70 Years*, Ontos, pages 420–455.
Sieg, W. (2008). 'Church without dogma: Axioms for computability'. In Cooper, S.B., Löwe, B.L., and Sorbi, A. (eds.), *New Computational Paradigms: Changing Conceptions of What is Computable*, Springer, pages 139–152.
Smith, P. (2007). *An Introduction to Gödel's Theorems*. Cambridge University Press.
Soare, R. I. (1986). *Recursively Enumerable Sets and Degrees*. Springer.
———. (1996). 'Computability and recursion'. *Bulletin of Symbolic Logic*, 2(3): 284–321.
———. (2009). 'Turing oracle machines, online computing, and three displacements in computability theory'. *Annals of Pure and Applied Logic*, 160(3): 368–399.
———. (2016). *Turing Computability: Theory and Applications*. Springer.
Sprigge, T. L. S. (1970). *Facts, Words and Beliefs*. Humanities Press.
Stalnaker, R. C. (1968). 'A theory of conditionals'. In Rescher, N. (ed.), *Studies in Logical Theory*, Blackwell, pages 98–112.
———. (1978). 'Assertion'. In *Context and Content: Essays on Intentionality in Speech and Thought* (1999), Oxford University Press, pages 78–95.
———. (1984). *Inquiry*. MIT Press.

———. (1996). 'Impossibilities'. In *Ways a World Might Be: Metaphysical and Anti-Metaphysical Essays* (2003), Oxford University Press, pages 55–67.
———. (2003). *Ways a World Might Be: Metaphysical and Anti-Metaphysical Essays*. Oxford University Press.
———. (2012). *Mere Possibilities: Metaphysical Foundations of Modal Semantics*. Princeton University Press.
Stalnaker, R. C. and Thomason, R. H. (1970). 'A semantic analysis of conditional logic'. *Theoria*, 36(1): 23–42.
Thomson, J. F. (1954). 'Tasks and super-tasks'. *Analysis*, 15(1): 1–13.
Todd, W. (1964). 'Counterfactual conditionals and the presuppositions of induction'. *Philosophy of Science*, 31(2): 101–110.
Turing, A. M. (1936). 'On computable numbers, with an application to the *Entscheidungsproblem*'. *Proceedings of the London Mathematical Society*, 42(1): 230–265.
———. (1939). 'Systems of logic based on ordinals'. *Proceedings of the London Mathematical Society*, 2(1): 161–228.
Welch, P. D. (2004). 'Post's and other problems of supertasks of higher type'. In Löwe, B., Piwinger, B., and Räsch, T. (eds.), *Classical and New Paradigms of Computation and Their Complexity Hierarchies* (Papers of the Conference "Foundations of the Formal Sciences III"), Kluwer Academic Publishers, pages 223–239.
Williamson, T. (2007). *The Philosophy of Philosophy*. Blackwell.
———. (2009). 'Probability and danger'. *The Amherst Lecture in Philosophy*, Lecture 4, pages 1–35.
———. (2010). 'Modal logic within counterfactual logic'. In Hale, B. and Hoffmann, A. (eds.), *Modality: Metaphysics, Logic, and Epistemology*, Oxford University Press, pages 81–96.
———. (2013). *Modal Logic as Metaphysics*. Oxford University Press.
———. (2015). 'Counterpossibles'. In Brochhagen, T., Roelofsen, F., and Theiler, N. (eds.), *Proceedings of the 20th Amsterdam Colloquium*, Institute for Logic, Language and Computation, pages 30–40.
Yablo, S. (2014). *Aboutness*. Princeton University Press.