The Paradox of Counterfactual Tolerance

Daniel Berntson

Counterfactuals are somewhat tolerant. Had Socrates been at least six feet tall, he need not have been *exactly* six feet tall. He might have been a little taller—he might have been six one or six two. But while he might have been a little taller, there are also limits to how tall he would have been. He would not have been a thousand feet tall, for example. Counterfactuals are not just tolerant, then, but bounded.

This paper presents a surprising paradox. If counterfactuals are both tolerant and bounded, then we can prove a flat contradiction using natural rules of inference. Moreover, not only are these rules of inference natural, they are also generally validated by our best semantic theories. These include the familiar Lewisian analysis in terms of similarity. Something has to go then. But what?

Putting my own cards on the table: I think that counterfactuals are both tolerant and bounded, and that this places an important *constraint* on counterfactual semantics. My own solution to the problem is to deny that similarity is transitive. This corresponds to analyzing counterfactuals in terms of *sufficient* instead of precise similarity. We will of course have much more to say about all this as we go along. But first, let's start with the paradox.

1. Paradox

Planck lengths are incredibly small. You would quite literally need a hundred million trillion of them just to span the diameter of a proton. Using Planck lengths as our basic unit of measure, we claim that counterfactuals are both tolerant and bounded.

Tolerance: For any h, had Socrates been at least h, he might have been at least $h + 1$.

Boundedness: There are j and k such that had Socrates been at least j, he would not have been at least k.

Tolerance says that had Socrates been at least six feet, he might have been at least a Planck length taller, and likewise for other heights. Boundedness says that there are heights j and k such that k is a **counterfactual bound** of j. So for example, it might be that had Socrates been at least six feet, he would not have been at least a thousand feet. In that case, a thousand feet will be a counterfactual bound of six feet, and boundedness will be satisfied.

Besides thinking that counterfactuals are tolerant and bounded, we also think that certain inferences preserve truth. For example, suppose that had the Athenians invaded Sparta, they would have lost and might have used catapults. It would then seem to follow that had they invaded Sparta *and* used catapults, they would have lost. This reasoning is perfectly natural and backed by an axiom called **rational monotonicity**.

$$\text{RM} \quad (A \mathrel{\Box\!\!\rightarrow} C) \wedge (A \mathrel{\Diamond\!\!\rightarrow} B) \supset (A \wedge B \mathrel{\Box\!\!\rightarrow} C)$$

As you can see, rational monotonicity is a kind of restricted strengthening rule. It tells us that we can strengthen from $A \mathrel{\Box\!\!\rightarrow} C$ to $A \wedge B \mathrel{\Box\!\!\rightarrow} C$ under certain special circumstances. Those special circumstances are that $A \mathrel{\Diamond\!\!\rightarrow} B$.[1]

We can now prove a flat contradiction. Below, h represents an open sentence saying that Socrates is at least h Planck lengths tall. Likewise for k and $n + 1$ and so on. Classical predicate logic is used throughout.

[1]As we go along, we will sometimes run together talk about *rules* and *conditional axioms*. The reason being that for present purposes, nothing much hangs on this distinction, and keeping track of it would be tedious. But we should point out that officially, RM is a conditional axiom. Given modus ponens, this makes it strictly stronger than the corresponding rule $A \mathrel{\Box\!\!\rightarrow} C, A \mathrel{\Diamond\!\!\rightarrow} B \vdash A \wedge B \mathrel{\Box\!\!\rightarrow} C$.

(1)	$j \mathrel{\Box\!\!\rightarrow} \neg k$	boundedness
(2)	$n \mathrel{\Box\!\!\rightarrow} \neg k$	hypothesis
(3)	$n \mathrel{\Diamond\!\!\rightarrow} n+1$	tolerance
(4)	$n \wedge (n+1) \mathrel{\Box\!\!\rightarrow} \neg k$	2, 3, RM
(5)	$n+1 \mathrel{\Box\!\!\rightarrow} \neg k$	4, substitution
(6)	$k-1 \mathrel{\Box\!\!\rightarrow} \neg k$	1, 2, 5, induction
(7)	$k-1 \mathrel{\Diamond\!\!\rightarrow} k$	tolerance
(8)	$\bot$	6, 7, duality

We have proved absurdity, so something has to go. Either one of our premises is false or one of our inference rules is invalid. Call this the **tolerance paradox**.[2]

2. Premises

Our paradoxical argument has two premises and, so, we might try to resolve the paradox by denying one or the other.

Suppose we deny boundedness. This means accepting that had Socrates been any taller, he might have been any height whatsoever. But any height whatsoever? We generally think that even if Socrates had been taller, the Spartans would not have thrown their weapons into the sea. But if Socrates might have been any height *whatsoever*, this is simply false. He might have been tall enough to crush Sparta with a single step. And in that case, the Spartans *would* have thrown their weapons into the sea. Who wants to oppose the Great Giant of Athens? If we deny that counterfactuals are bounded, so many ordinary counterfactuals will turn out false that we might as well give up counterfactual reasoning altogether.

Two more things should be pointed out about boundedness. First, boundedness says that some height has a counterfactual bound. It does not say that *all* heights have a counterfactual bound. Sufficiently outlandish heights might still be unbounded. For all boundedness cares, it may be that had Socrates been at least a hundred million trillion feet tall, he might

[2]

have been any height whatsoever. Even so, boundedness will still hold if there is some pair of heights j and k such that had Socrates been at least j, he would not have been at least k. Six feet and a thousand feet would seem to do the trick. Even if Socrates might have been any height whatsoever, had he been at least a hundred million trillion feet, he would not have been at least a thousand feet, had he been at least six feet.

Second, boundedness does not require there to be a height with a *least* counterfactual bound. If it did, there might be concerns about sharp cutoffs and whether or not boundedness was determinately true. But since boundedness only requires there to be a height with *some* counterfactual bound, those concerns can be sidestepped. It may be vague whether or not Socrates might have been at least twelve feet, had he been at least six feet. But even if so, it is *not* vague whether or not he might have been a thousand feet. He determinately would not have been.

Suppose that instead of denying boundedness, we deny tolerance. This means accepting the existence of what we might call **singularities**. Singularities are just heights such that had Socrates been at least that tall, he would not have been even a Planck length taller.

As it turns out, there are certain good reasons to accept the existence of singularities. One of them is **strong centering**.

SC $\quad (A \wedge B) \supset A \mathrel{\Box\!\!\rightarrow} B$

This says that whenever A and B are both true, then $A \mathrel{\Box\!\!\rightarrow} B$ is also true. So for example, say that Socrates is exactly five feet tall. It then follows that he is at least five feet and that he is not at least a Planck length taller. So by strong centering, had Socrates been at least five feet, he would not have been even a Planck length taller. The actual height of Socrates thus turns out to be a singularity, given strong centering.

Another reason to think that there are singularities is that certain heights may be impossible. For example, it may be biologically possible for Socrates to have been up to fifteen feet tall, but biologically impossible for him to have been any taller. In that case, fifteen feet would seem to be a singularity, and so a counterexample to tolerance.

Why is that? There are certain plausible connections between

possibility, necessity, and counterfactuals. For example, it would seem that for any genuine notion of possibility and corresponding notion of necessity, a would counterfactual is true whenever the antecedent is possible and the consequent is necessary. That is:

$$\text{M} \quad (\Diamond A \wedge \Box B) \supset A \mathrel{\Box\!\!\rightarrow} B$$

This rule correctly predicts that since Socrates is possibly a farmer and necessarily human, it follow that had Socrates been a farmer, he would have been human.[3] This is exactly the sort of thing you would expect.

Now suppose that biological possibility is a genuine form of possibility. And suppose that it is biologically possible for Socrates to have been fifteen feet tall, but biologically impossible for him to have been any taller. This means that it is biologically possible for Socrates to have been at least fifteen feet and biologically necessary that he is not at least fifteen feet plus a Planck length. Given the above inference rule, it follows that had Socrates been at least fifteen feet, he would not have been at least fifteen feet plus a Planck length. So fifteen feet is a singularity and an apparent counterexample to tolerance.

Special circumstances can also lead to singularities. Suppose that Socrates is in fact genetically engineered by a team of alien scientists. Their advanced technology lets them precisely control his height and, as it turns out, their research grant requires them to make his height a prime number of Planck lengths. But research funding being what it is, the scientists also have a strong preference for smaller heights, since smaller humans are cheaper to build. The result is that for every prime height greater than his actual height, had Socrates been at least that tall, he would have been exactly that tall. All those heights are singularities.

There are different ways of dealing with such difficulties. We could deny strong centering. We could deny the proposed link between

[3]Given that the necessity and possibility operators are duals ($\Diamond A \equiv \neg\Box\neg A$) and that the would and might counterfactual operators are duals ($A \mathrel{\Diamond\!\!\rightarrow} B \equiv \neg(A \mathrel{\Box\!\!\rightarrow} \neg B)$, this rule is equivalent to the rule $\Diamond A, A \mathrel{\Diamond\!\!\rightarrow} B \vdash \Diamond B$. That rule is discussed at some length in **?** 64-67.

possibility, necessity, and would counterfactuals. The simplest solution, though, is to just restrict tolerance and boundedness. What we claim is that there is some *range* of heights R such that for every height $h \in R$, had Socrates been at least h, he might have been at least $h + 1$. And for some pair of heights $j, k \in R$, had Socrates been at least j, he would not have been at least k. This is all we need to run the paradox. The original way we formulated tolerance and boundedness, while simpler, is somewhat stronger than what we actually need.

Moreover, that there are such ranges of heights would seem to be obvious. For example, take the range of heights between six feet and nine feet. Had Socrates been at least six feet, he would not have been at least nine feet. And for every height between six feet and nine feet, had Socrates been at least that tall, he might have been a Planck length taller. Strong centering is no longer a problem, since Socrates is actually five feet, and five feet is well below the bottom of the range. And the impossibility of Socrates being taller than fifteen feet is no longer a problem, since fifteen feet is well above the top of the range.

Putting the issue in the most general terms: The rules of counterfactual inference are necessarily valid, if valid at all. So all we need is for it to be *possible* for there to be *some* range of *some* quantity that satisfies both tolerance and boundedness. If so, then we can reason to paradox.

3. Auxiliary Inferences

The tolerance paradox uses various auxiliary inference rules and, so, you might wonder if they are the source of the problem. These include classical predicate logic and mathematical induction, which are not promising points of resistance. Giving up either means giving up modern mathematics, and we are not going to give up modern mathematics. The cost is simply too high.

Moreover, even if we *were* willing to give up induction or classical logic, it is not especially clear how that would help. Our paradoxical proof is intuitionistically valid, so valid by the lights of the most promising alternative to classical logic. And while induction is convenient, it is also

not really essential. Suppose that we agree that some particular height has a certain counterfactual bound. Suppose we agree that had Socrates been at least six feet, he would not have been at least nine feet. We could then show that had Socrates been at least a thousand feet, he would not have been even a Planck length taller using a very long—but still finite—proof. The proof would use the sam basic reasoning, but not require induction.

This leaves two remaining auxiliary rules. The first is **substitution**.[4] Substitution says that when two propositions are equivalent, we can replace the one with the other in the antecedent of a counterfactual. Equivalent in what sense? There are several options here. My own inclination is to use counterfactual equivalence.[5] But the relevant propositions are also necessarily equivalent, a priori equivalent, analytically equivalent, and maybe even logically equivalent. Basically, *any* reasonable rule telling us that we can substitute equivalents will let us replace the proposition that Socrates is at least n and at least $n+1$ with the proposition that he is at least $n+1$.

Moreover, even if all such substitution principles were to fail, the inference would still go through by other means. Thinking of substitution as substitution under counterfactual equivalence and making the counterfactual equivalence explicit, the original reasoning looks like this:

(4)	$n \wedge (n+1) \boxrightarrow \neg k$	2, 3, rational monotonicity
(5)	$(n \wedge (n+1) \boxrightarrow n) \wedge (n \boxrightarrow n \wedge (n+1))$	nature of heights
(6)	$n+1 \boxrightarrow \neg k$	4,5, substitution

But even without substitution, we can get the same effect using other rules. For example, we could use **limited transitivity** to reason as follows.[6]

[4] $A \boxrightarrow C \supset B \boxrightarrow C$ when A and B are equivalent.

[5] A and B are counterfactually equivalent when $A \boxrightarrow B$ and $B \boxrightarrow A$.

[6] Limited transitivity tells us is that whenever we have a would counterfactual with conjunctive antecedent and one of the conjuncts counterfactually entails the other, the conjunct that is counterfactually entailed can be eliminated. In the present case, it is probably best thought of as a pair of axioms:

(4)	$n \wedge (n+1) \mathrel{\Box\!\!\rightarrow} \neg k$	2, 3, rational monotonicity
(5)	$n+1 \mathrel{\Box\!\!\rightarrow} n$	nature of heights
(6)	$n+1 \mathrel{\Box\!\!\rightarrow} \neg k$	4,5, limited transitivity

The rest of the proof then goes through as before. This means that denying substitution—on its own at least—will not resolve the paradox.

The other auxiliary inference rule is **duality**.[7] Duality says there is a certain equivalence between might and would counterfactuals. This is widely accepted, but also controversial. Fortunately, while posing the paradox using duality is natural, it is also not required. One reason is that in the original argument, we used induction and duality to reason as follows:

(6)	$k-1 \mathrel{\Box\!\!\rightarrow} \neg k$	1, 2, 5, induction
(7)	$k-1 \mathrel{\Diamond\!\!\rightarrow} k$	tolerance
(8)	$\bot$	6, 7, duality

But even without duality, we could still have used induction to directly inferred that:

(6)	$k \mathrel{\Box\!\!\rightarrow} \neg k$	1, 2, 5, induction

This is bad enough. Thinking of k as a thousand feet, this says that had Socrates been at least a thousand feet, he would *not* have been at least a thousand feet. This strikes me as not just false, but inconsistent.[8] So we

---

$(A \mathrel{\Box\!\!\rightarrow} B) \wedge (A \wedge B \mathrel{\Box\!\!\rightarrow} C) \supset A \mathrel{\Box\!\!\rightarrow} C$

$(B \mathrel{\Box\!\!\rightarrow} A) \wedge (A \wedge B \mathrel{\Box\!\!\rightarrow} C) \supset B \mathrel{\Box\!\!\rightarrow} C$

The reason that we need two axioms is that we need one for each conjunct. Of course, if we had *substitution*, then only the first would be needed, since the antecedents are logically equivalent. But the status of substitution is currently in question, so we need two.

[7] $A \mathrel{\Box\!\!\rightarrow} B \equiv \neg(A \mathrel{\Diamond\!\!\rightarrow} \neg B)$

[8] Why think that $\vdash \neg(k \mathrel{\Box\!\!\rightarrow} \neg k)$? First, we claim that $\vdash (\Diamond A) \wedge (A \mathrel{\Box\!\!\rightarrow} B) \supset \Diamond(A \wedge B)$ for any genuine notion of possibility and necessity. But now take the case in which $\Diamond$ and $\Box$ denote *logical* possibility and necessity respectively. Since $\nvdash \neg k$ and $\vdash \neg(k \wedge \neg k)$, we thus have $\vdash \Diamond k$ and $\vdash \neg\Diamond(k \wedge \neg k)$. But then since logical possibility and necessity are genuine forms of possibility and necessity, it follows that $\vdash \neg(k \mathrel{\Box\!\!\rightarrow} \neg k)$ by the proposed principle and propositional logic.

still have paradox. Denying duality is no help.[9]

4. Rational Monotonicity

Rational monotonicity is natural and part of counterfactual common sense. Had the Athenians invaded Sparta, they would have lost and might have used catapults. What could be more natural than concluding that had they invaded *and* used catapults, they would have lost?

Besides being natural, rational monotonicity is also generally validated by our best counterfactual semantics. A prominent example is the similarity semantics from **?**. He analyzes counterfactual using what we are going to call **similarity models**. These are triples:

$\langle W, R, v \rangle$

where W is a set of worlds v is a valuation function. R is a three-place counterfactual accessibility relation on worlds. This means that, strictly speaking, R is just a set of triples.[10] But Lewis think of it as a *similarity* relation on worlds, so we can too.

$Rwvu$ w is no less similar than v to u

Because Lewis thinks of R as strict similarity, he requires it to be total and transitive and so form a weak total ordering.

Total	either $Rwvu$ or $Rvwu$
Transitive	if $Rwvz$ and $Rvuz$, then $Rwuz$

It would seem to be clear that the similarity relation has these features. Suppose we are comparing similarity relative to some fixed world z. Totality then says that there are never pairs of worlds such that each is less similar than the other. Transitivity says that if w is no less similar than v and v is no less similar than u, then w is no less similar than u.

[9] The other reason that denying duality is no help is that we can reformulate the original paradox using only might counterfactuals. That version of the argument can be found in section 11.

[10] That is, $R \subset W^3$.

Once we have our strict similarity models, we can define the counterfactual operators.

$u \models A \boxrightarrow B$ iff there is a world $v \in W$ such that $v \models A$ and every world $w \in W$ is such that $w \models A \supset B$ whenever $Rwvu$

$u \models A \diamond\!\!\rightarrow B$ iff for every world $v \in W$ such that $v \models A$, there is a world $w \in W$ is such that $w \models A \wedge \neg B$ and $Rwvu$

To simplify matters a bit, suppose we have the **limit assumption** in place.[11] What these clauses are then telling us is that:

$A \boxrightarrow B$ is true iff all of the most similar A worlds are B worlds.

$A \diamond\!\!\rightarrow B$ is true iff some of the most similar A worlds are B worlds.

Here, an A world is among the most similar A worlds when it is no less similar than any other A world.

But now here is something we can prove. We can prove that rational monotonicity is valid over the class of all strict similarity models.[12] This means that if we are going to resolve the counterfactual tolerance paradox by denying rational monotonicity, we will have to reject the Lewisian analysis of counterfactual in terms of similarity.

[11] The limit assumptions says that there are no infinite descending chains of increasing similarity. More formally, for every $X \supset W$ and any $u \in W$, there is a $w \in X$ such that $Rwvu$ for every $v \in X$.

[12] We need to show that $\models (A \boxrightarrow C) \wedge (A \diamond\!\!\rightarrow B) \supset (A \wedge B \boxrightarrow C)$. Suppose then that $A \boxrightarrow C$ and $A \diamond\!\!\rightarrow B$ are both true at a world z. Given the semantics, this means that there is a world v at which A such that all of the worlds w that are at least as similar to z are worlds at which $A \supset C$ and that one of those worlds u is a world at which $A \wedge B$. Now to show that $A \wedge B \boxrightarrow C$ is true at z, all we need to show is that every world u^* that is at least as close to z as u is a world at which $A \wedge B \supset C$. To show that, suppose otherwise. Suppose there is a world u^* that is at least as close to z as u and also a world at which $A \wedge B \wedge \neg C$. Then since worlds are logically closed, u^* is a world at which $A \wedge \neg C$. But then by transitivity, since u^* is at least as close as u to z and u is at least as close as v to z, that u^* is at least as close as v to z. But then $A \boxrightarrow C$ is false, contrary to assumption. So $A \wedge B \boxrightarrow C$ is true at z. Therefore $\models (A \boxrightarrow C) \wedge (A \diamond\!\!\rightarrow B) \supset (A \wedge B \boxrightarrow C)$.

5. Uniform Intolerance

Now to my knowledge, Lewis has no official position on the tolerance paradox, but his earlier response to Pollock is telling. Lewis would rather give up infinite agglomeration than give up the idea that for every length longer than an inch, had the line been longer, it would have been shorter than that. This suggests that not only would Lewis deny that counterfactuals are tolerant, he would do so with characteristic gusto. He would claim that counterfactuals are **uniformly intolerant**. Letting g be the actual height of Socrates, Lewis would say that for every $h > g$, had Socrates been at least h, he would have been exactly h. This because any world in which Socrates is taller than h is less similar to our world than a world in which he is exactly h. It's not just that some of the heights greater than g are singularities. All of them are. This follows from the analysis of counterfactuals in terms of similarity.

There are certain advantages to uniform intolerance. In comparison to coarse graining, it removes any need to explain why some heights are singularities and others are not. But while that may be, uniform intolerance runs roughshod over how we ordinarily think about counterfactuals.

Suppose for example that Nixon has a finicky nuclear button. To prevent accidental launches, the button has to be pressed at least n Planck lengths to send a signal. But even when pressed at least $n + m$ Planck lengths, there is no guarantee that it launches a rocket. If m is even, the button sends a weak signal that fizzles out before reaching the launch pad. If m is odd, the button sends a strong signal that launches a rocket. As it happens, Nixon presses the button, but not quite hard enough to send a signal. What would have happened if Nixon had pressed a little harder? Uniform intolerance tells us that everything would have been fine. Had Nixon pressed the button at least n Planck lengths, he would have pressed it exactly n Planck lengths. A weak signal would have been sent and it would have fizzled out. But this is the wrong result: Had Nixon pressed hard enough to send signal, he might have pressed it a little past n. He might have launched a rocket and might have started a war.

We know that Nixon might have launched a rocket, in part, because

of the role counterfactuals play in justifying our emotions. We should be *relieved* that Nixon did not push the button hard enough to send a signal. But if we know that had Nixon pushed the button, he would not have launched a rocket, why are we relieved? Nothing bad would have happened. Or suppose that Kissinger offers Nixon a dollar to push the button lightly and two dollars to push it hard enough to send a signal. Nixon pushes it lightly to get the dollar, but then reflects: Had he pressed the button hard enough to send a signal, he would not have a launched a rocket, but would have gotten an extra dollar. He thus *regrets* not pushing the button. All he did was leave money on the table. But this is obviously silly. Nixon should have no regrets because, had he pushed the button hard enough, he would have gotten an extra dollar, but he also might have started a nuclear war. He did the right thing.

Putting the issue more generally: Lewis has built his semantics to avoid certain kinds of arbitrary choices. Had you flipped a fair coin, it might have landed heads and it might have landed tails. It is simply false that it *would* have landed heads. If it were true that it would have landed heads, then reality would have to somehow arbitrarily choose the heads scenario over the tails scenario, despite the fact that those scenarios are symmetric and so equally similar.[13] The problem with uniform intolerance is that reality will have to make choices that are *almost* completely arbitrarily. Perhaps the world where Nixon pushes the button n Planck lengths is marginally more similar to our world than the one where he pushes it $n + 1$ Planck lengths. But the choice between these worlds is still mostly arbitrary. The choice between is just about as unmotivated as the choice between the world where you flip heads the world where you flip tails. You would think, then, that this is the sort of choice Lewis would want to avoid.

[13] Stalnaker of course has a line of response to this sort of reasoning. We will say more about his selection models in section 11

6. Vagueness

At this point, you might be convinced that we have a paradox, but concerned that we do not have a new paradox. Why is this not just the sorites paradox in another form?

To build a sorites paradox, we need a scale and something like a vague all-or-nothing predicate. We might observe, for example, that one grain of sand is not a heap, but that a thousand grains of sand is a heap. The scale is the number of grains of sand. The vague predicate is being a heap. We then reason as follows:

(1)	One grain is not a heap.	premise
(2)	A thousand grains are a heap.	premise
(3)	It is not the case that (n grains are not a heap and $n+1$ grains are a heap).	premise
(4)	If n grains are not a heap, then $n+1$ grains are not a heap.	3, PL
(5)	A thousand grains are not a heap.	1,4, induction
(6)	$\bot$	2,5, PL

As you can see, the paradox gets going because we deny that there are any sharp cutoffs. We deny that there is any number of grains n such that n is not a heap, but that $n+1$ is a heap. But then given the denial of sharp cutoffs, we can reason to a flat contradiction using induction and classical logic.

Now in the case of counterfactuals, we certainly *can* build a sorites paradox. We can use heights as the scale and the counterfactual operators in place of a vague all-or-nothing predicate. Like the original sorites argument, that argument depends on the denial of sharp cutoffs. What we deny in the case of counterfactuals that there is any sharp cutoff in how tall Socrates might have been.

> **No Sharp Cutoffs:** There are no heights m and n such that (a) it is true that had Socrates been at least m, he might have been at least n and (b) it is false that had Socrates been at least m, he might have been at least $n+1$.

Besides denying that there are sharp cutoffs, we also have a boundedness condition. We claim that:

> **Boundedness:** There are heights m and k such that $m < k$ and it is not the case that had Socrates been at least m, he might have been at least k.

The sorites paradox for counterfactuals then runs as follows:

(1)	$m \diamondsuit\!\!\rightarrow m$	theorem
(2)	$\neg(m \diamondsuit\!\!\rightarrow k)$	boundedness
(3)	$\neg(m \diamondsuit\!\!\rightarrow n \wedge \neg(m \diamondsuit\!\!\rightarrow n+1))$	no sharp cutoffs
(4)	$(m \diamondsuit\!\!\rightarrow n) \supset (m \diamondsuit\!\!\rightarrow n+1)$	3, PL
(5)	$m \diamondsuit\!\!\rightarrow k$	1,4 induction
(6)	$\bot$	2,5, PL

Think of m as six feet and k as a thousand feet. The paradoxical argument thus starts with the observation that had Socrates been at least six feet, he might have been at least six feet. This will be a theorem of any reasonable counterfactual logic. We then deny that had Socrates been at least six feet, he might have been at least a thousand feet and deny that there are is a sharp cutoffs with respect to how tall Socrates might have been, had he been at least six feet tall. But then by classical logic and induction, we get a flat contradiction.

The key observation is that while the sorites paradox is certainly a paradox, the sorites paradox is not the tolerance paradox. The two paradox are distinct. They have different premises and rely on different inference rules.

Start with the premises. Both paradoxes rely on a kind of boundedness condition, so they have that much in common. The difference is that where the tolerance paradox depends on the acceptance of tolerance, the sorites paradox depends on the denial of sharp cutoffs. These are clearly different claims. On the one hand, the denial of tolerance entails the acceptance of sharp cutoffs. After all, if we deny tolerance, then there is some m such that:

$$\neg(m \diamondsuit\!\!\rightarrow (m+1))$$

But from this it follows that:

$$m \diamondsuit\!\!\rightarrow m \wedge \neg(m \diamondsuit\!\!\rightarrow (m+1))$$

So there is some n such that:

$$m \diamondsuit\!\!\rightarrow n \wedge \neg(m \diamondsuit\!\!\rightarrow \neg(n+1))$$

That is, there is a sharp cutoffs with respect to how tall Socrates might have been, had he been at least m tall. On the other hand, going the other direction, the acceptance of tolerance *does not* entail the denial of sharp cutoffs. You might, for example, accept that for every m:

$$(m \diamondsuit\!\!\rightarrow m+1) \wedge \neg(m \diamondsuit\!\!\rightarrow m+2)$$

Had Socrates been at least six feet tall, he might have been at least one Planck length taller, but would not have been at least two Planck lengths taller, and likewise for other heights. But in that case, we have both tolerance and sharp cutoffs. The acceptance of tolerance follows from the first conjunct. Letting $n = m+1$, the acceptance of sharp cutoffs also follows, since:

$$(m \diamondsuit\!\!\rightarrow n) \wedge \neg(m \diamondsuit\!\!\rightarrow n+1)$$

In that case, the tolerance paradox will go through, but the sorites paradox will not. So the two paradoxes are not the same. The acceptance of tolerance is *strictly weaker* than the denial of sharp cutoffs.

Another reason to think that the paradoxes are distinct is that they rely on different inference rules. If you look back at the sorites paradox for counterfactuals, we used counterfactual logic to get the theorem on the first line, which was the claim that had Socrates been at least m, he might have been at least m. But we could have just taken that as a premise and, in that case, we would have had a paradox without using *any* counterfactual inferences rules. This stands in stark contrast to the tolerance paradox, which relies on counterfactual inference rules like rational monotonicity.

Going the other way, the sorites paradox requires inference rules that the tolerance paradox does not. In particular, the sorites paradox requires us to infer the material conditional on line four from the denial of sharp

cutoffs on line three. But while that inference is classically valid, it is not intuitionistically valid—it is not valid if we go along with the intuitionists and deny the law of the excluded middle. The tolerance paradox, on the other hand, is intuitionistically valid. It goes through even if we deny the law of the excluded middle. So this is yet further reason to think that the paradoxes are genuinely distinct.

7. Lines

The tolerance paradox has a certain resemblance to a paradox from John **?**. Suppose you draw a one inch line on a piece of paper. We then consider, how long would the line have been, had it been longer than an inch? Well, a world in which the line is less than two inches is more similar than one in which it is two inches. The familiar Lewisian similarity semantics then tells us that had the line been longer than an inch, it would have been shorter than two inches. But then if space is continuous, the same goes for all the other lengths between two inches and one inch. Given **infinite agglomeration**, it then follows that had the line been longer than an inch, it would *not* have been longer than an inch.[14] So we have paradox.

There are important differences between our paradox and Pollock's line paradox. The first is that his paradox assumes that we are using something like Lewisian similarity semantics. Otherwise, there is no way to infer that (a) had the line been longer than an inch, it would have been shorter than two inches from (b) the fact that worlds in which the line is less than two inches are more similar to ours than worlds in which it is two inches. Our paradox, on the other hand, makes no assumptions about semantics. All we assumed was that counterfactuals were both tolerant and bounded and that certain rules of inference were valid. Now as a matter of fact, Lewisian similarity semantics validates the relevant rules, and so is subject to our paradox. But *any* semantics that validated those rules would be in the same position. There is no need to assume any particular connection between counterfactuals and similarity.

[14] $\wedge_i(A \mathrel{\Box\!\!\rightarrow} B_i) \supset (A \mathrel{\Box\!\!\rightarrow} \wedge_i B_i)$

Second, the line paradox not only depends on using something like Lewisian similarity semantics. It depends on using what we might call offhand similarity as the similarity relation on worlds. Offhand, a line that is one and a half inches long is more similar to a line that is one inch long than a line that is two inches long. It must follow, then, that in the sense that matters for counterfactuals, a *world* in which a line is one and a half inches long is more similar to our world than a world in which it is two inches long.

To his credit, Pollock recognizes that this is a substantial assumption and, in fact, suggests that we could resolve his paradox by denying it. Rather than using offhand similarity for our counterfactual semantics, we could use a specialized **coarse grained** similarity relation, one that counts worlds in which the line is between one inch and two inches, say, as equally similar. We then no longer get the result that had the line been longer than an inch, it would not have been longer than an inch.

There is much to be said for using a specialized similarity relation when doing similarity semantics. The important point for our purposes, though, is that while a coarse grained similarity relation might solve the line paradox, it gets us nowhere with the tolerance paradox.

Suppose that we use a similarity ordering that counts all the worlds in which Socrates is between six feet and seven feet as equally similar. This lets us make sense of the idea that had Socrates been at least six feet, he might have been at least a Planck length taller. So far so good. The problem now is the worlds at the top of the equivalence class—those worlds at which Socrates is exactly seven feet tall. All the worlds in which he is even a Planck length taller are in the next similarity class. So had Socrates been at least seven feet, he would not have been even a Planck length taller. We thus get a failure of tolerance. Coarse graining solves the problem only if we forget what the problem was in the first place.

The third difference between the line paradox and ours is that his paradox relies on infinite agglomeration. One way of resolving his paradox, then, is to deny its validity. **?** does exactly that: He accepts that for every length longer than an inch, had the line been longer, it would have been shorter than that. What he denies is that it follows that had

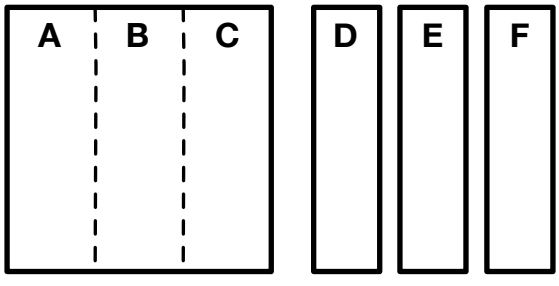
the line been longer than an inch, it would not have been longer than an inch. The tolerance paradox, on the other hand, has nothing to do with agglomeration, nor does it use any infinitary inference rules. So the tolerance paradox and the line paradox are clearly quite different.

8. Woody

The tolerance paradox, as we have seen, is quite different from both the sorites paradox and the line paradox. What I want to suggest in this section is that there is in fact a much better analogy. The tolerance paradox< it seems to me, is a kind of counterfactual analogue of a certain modal paradox from **?** and **?**.

Consider a particular table named Woody. Ordinarily speaking, we think that things like tables could have been built out of slightly different matter. Had a few of the particles that compose Woody not existed, Woody could still have existed. His existence may depend on his matter, but it does not depend on *all* his matter. The composition of Woody is somewhat tolerant, we might say. That said, there are also limits to how different his matter could have been. Woody could have been built out of slightly different matter, sure. But he could not have been built out of *entirely* different matter.

Making all of this somewhat more precisely, think of Woody as a tabletop rather than a table, one composed of three planks *ABC*. There are then three spare planks *D*, *E*, and *F* set to the side. Like this:



We then claim that Woody's composition is both modally tolerant and modally tolerant in the following sense.

Modal Tolerance: Necessarily, if Woody is made out of three planks, then he could have been made out of only two of those planks, together with some other third plank.

Modal Boundedness: Necessarily, if Woody is made out of three planks, then he could not have been made out of three entirely different planks.

So for example, Woody is in fact made out of *ABC*. Modal tolerance then tells us that he could have been made out of *BCD*, since being made out of *BCD* means that Woody would have been made out of two actual planks and one spare plank. Modal boundedness says that he could not have been made out of *DEF*, since he would have been made out of entirely different planks.

Besides thinking that composition is tolerant and bounded, we also think that certain modal inferences preserve truth. For example, it would seem that:

S4 $\quad \Diamond\Diamond A \supset \Diamond A$

Suppose that it could have been that Socrates was a possible fishmonger. In that case, it would seem to follow that Socrates is in fact a possible fishmonger. The possible possibilities are also possibility, or so we generally think. Or equivalently, since possibility is the dual of necessity, we generally think that the necessities are themselves necessary. If Socrates is necessarily a human, then it is *necessary* that Socrates is necessarily a human. The propositions that are necessary is not itself a contingent matter.

Together, of course, these convictions lead to disaster. Because if constitution is modally tolerant and modally bounded and S4 is valid, we can prove a flat contradiction.

(1)	ABC	premise
(2)	$\neg\Diamond DEF$	modal boundedness
(3)	$\Box(ABC \supset \Diamond BCD)$	modal tolerance
(4)	$\Diamond BCD$	1,3, T
(5)	$\Box(BCD \supset \Diamond CDE)$	modal tolerance
(6)	$\Diamond\Diamond CDE$	4,5, K
(7)	$\Diamond CDE$	6, S4
(8)	$\Box(CDE \supset \Diamond DEF)$	modal tolerance
(9)	$\Diamond\Diamond DEF$	7,8, K
(10)	$\Diamond DEF$	9, S4
(11)	$\bot$	2, 10, PL

Besides S4, the paradoxical proof makes use of classical logic and a pair of auxiliary modal axioms called T and K. But these are completely uncontroversial.[15] It looks, then, like we face a choice. We can deny modal tolerance, modal boundedness, or S4. That choice looks quite a bit like the choice we face in the counterfactual tolerance paradox.

There is of course an enormous literature on the Woody paradox and this is not the place to sift through all the possible responses. My own preferred response, though, is to deny S4. And how this is done in the modal case, it seems to me, suggests a way forward in the counterfactual case. This will be the topic of the next section.

9. Transitivity and Counterfactuals

Most readers will be familiar with Kripke models for modal logic. These are triples, much like the similarity models we saw in section 4.

$$\langle W, R, v\rangle$$

The main difference is that where the counterfactual accessibility relation is a three-place relation, the modal accessibility relation is just a two-place

[15] T says that $\vdash \Box A \supset A$ and K says that $\vdash \Box(A \supset B) \supset \Box A \supset \Box B$. The first axiom corresponds to the idea that necessary truths are *truths*. The second, more or less, to the idea that a proposition is necessary iff it is true at all possible worlds.

relation. But otherwise, the basic idea is surprisingly similar. We can then analyze the necessity and possibility operators in the usual way:

$w \models \Box A$ iff A is true at every $v \in W$ such that Rwv
$w \models \Diamond A$ iff A is true at some $v \in W$ such that Rwv

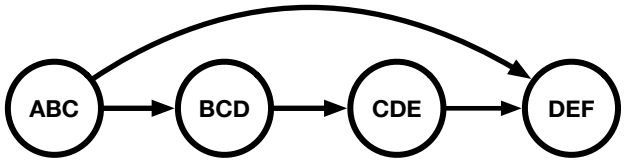
A sentence is necessary at a world when it is true at all accessible worlds. A sentences is possible at a worlds when it is true at some accessible worlds.

The observation now that there is a certain connection between the modal axiom S4 and the transitivity of the modal accessibility relation.

Transitivity: If Rwv and Rvu, then Rwu

In particular, the modal axiom is valid iff we require modal accessibility relation to be transitive.[16]

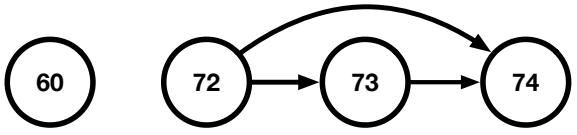
You can see the basic problem that transitivity raises. Suppose, as before, that we endorse both modal tolerance and modal boundedness. Given transitivity, modal space then ends up looking like this:



The actual world is one in which Woody is built out of ABC. Because he could have been built out of only two-thirds of his matter, he could have been made out of BCD. So BCD is accessible from ABC. But it is true at ABC that Woody *necessarily* could have been made out of two-thirds different matter. So CDE is accessible from BCD. Transitivity then tells us that it follows that CDE is accessible from ABC. But now, since CDE is accessible from ABC and it is necessarily possible at ABC that Woody is made out of only two-third of his matter, DEF is accessible from CDE. But then by transitivity again, DEF is accessible from ABC. So it is true at ABC that Woody could have been made out of entirely different matter. But that contradicts modal boundedness.

[16] Or more precisely, it is valid over precisely the class of transitive frames.

It turns out that transitivity cause more or less *the same problem* in the case of counterfactuals. For simplicity, think of counterfactual tolerance in terms of inches instead of Planck lengths. What we claim is that for any height great than his actual height, had Socrates been at least that tall, he might have been at least one inch taller. Now suppose that we think about things in terms of similarity models. We then get the following:



Suppose that the actual world is on the left and that all of our similarity comparison are fixed relative to the actual world. In that case, since it is true that had Socrates been at least 72 inches, he might have been 73 inches, it follows by our semantics that the 73 inch world is at least as similar as the 72 inch world. We represent that fact here using an arrow from the 72 inch world to the 73 inch world. But now by tolerance again, had Socrates been at least 73 inches tall, he might have been at least 74 inches tall. So the 74 inch world is at least as similar as the 73 inch world. But now by the *transitivity* of the similarity relation, the 74 inch world is at least as similar as the 72 inch world. But then by parity of reasoning, we can show that worlds in which Socrates is *arbitrarily* tall are at least as similar as the worlds in which he is 72 inches tall. But they by the Lewisian analysis of the might counterfactual, had Socrates been at least 72 inches, he might have been arbitrarily tall. So we get a contradiction of boundedness.

The problem, it would seem, is that *the similarity relation is transitive*. And so the solution would seem to be just as clear. We should keep the basic idea behind Lewisian similar similarity models, but give up transitivity.

10. Sufficient Similarity

What we are going to call **sufficient similarity models** are triples consisting of a set of worlds, a three-place accessibility relation on worlds, and a

valuation function.

$$\langle W, R, v \rangle$$

These are models are in most respects just like the familiar Lewisian similarities models from section 4. The difference is that compared to Lewis, we are placing fewer requirements on the counterfactual accessibility relation. In particular, instead of requiring it to be total and *transitive*, we require it to be total and (weakly) acyclic.

Total	Either $Rwvu$ or $Rvwu$
Acyclic	If $\neg Rwvz$ and $\neg Rvuz$, then $\neg Rwuz$

As before, the counterfactual accessibility relation is just a set of triples, so could in theory represent any three-place relation between worlds you want, so long as it has the right formal properties. But one natural way to think about is as representing facts about **sufficiently similarity**.

$Rwvu$	w is not sufficiently less similar than v to u

To help distinguish this relation from Lewis's relation, we can call his relation **precise similarity**.

You can see that sufficient similarity is total and acyclic. That it is total means that there are never pairs of worlds w and v such that each is sufficiently less similar than the other. That it is acyclic means that if w is sufficiently less similar than v and v is sufficiently less similar than u, then w is sufficiently less similar than u. But now notice that while sufficient similarity is total and acyclic, it is not generally transitive. So we have a relation of the right form.

Once we have our sufficient similarity models, the counterfactual operators are defined the same way as before. We can just reuse the Lewisian clauses from section 4.

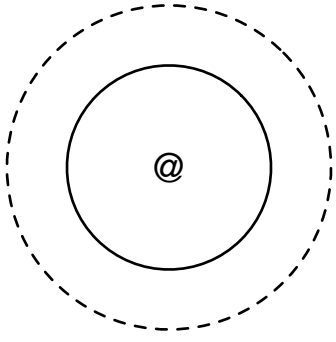
$u \models A \mathrel{\Box\!\!\rightarrow} B$	iff there is a world $v \in W$ such that $v \models A$ and every world $w \in W$ is such that $w \models A \supset B$ whenever $Rwvu$
$u \models A \mathrel{\Diamond\!\!\rightarrow} B$	iff for every world $v \in W$ such that $v \models A$, there is a world $w \in W$ is such that $w \models A \wedge \neg B$ and $Rwvu$

To simplify, suppose again that we have the limit assumption in place. What these clauses are then telling us is that:

> $A \mathrel{\Box\!\!\rightarrow} B$ is true iff all of the sufficiently similar A worlds are B worlds.
>
> $A \mathrel{\Diamond\!\!\rightarrow} B$ is true iff some of the sufficiently similar A worlds are B worlds.

Here, the sufficiently similar A worlds are just the A worlds that are not sufficiently less similar than any other A world

Perhaps somewhat more intuitively, here is the contrast with Lewis. Lewis says that $A \mathrel{\Box\!\!\rightarrow} B$ is true iff all of the most similar A worlds are B worlds. Below, the most similar A worlds are represented by the solid sphere around the actual world.



To determine whether or not $A \mathrel{\Box\!\!\rightarrow} B$, then, we have to check to see if the solid sphere includes any A worlds that are also $\neg B$ worlds. The sufficient similarity approach, on the other hand, says that $A \mathrel{\Box\!\!\rightarrow} B$ is true iff all of the *sufficiently* similar A worlds are B worlds. Here, the sufficiently similar A worlds are represented by the dotted sphere. As you can see, they include all of the most similar A worlds, but they also include A worlds that are just a bit less similar. So the region we need to check extends just a bit farther out into modal space.

Sufficient similarity models let us confirm that transitivity and rational monotonicity are in fact intimately related. That is, we can show that:

Rational monotonicity is valid on precisely the class of transitive frames.

Thinking in terms of similarity, denying the validity of rational monotonicity is equivalent to denying the transitive of the counterfactual accessibility relation.[17] This confirms the suggested analogy with the Woody paradox. Rational monotonicity is counterfactual logic as S4 is to modal logic. Both principles are valid when the accessibility relation is required to be transitive. And both principles fail once transitivity is denied.

11. Selection

The perhaps most prominent competitor to Lewisian semantics is the selection semantics from Robert Stalnaker.[18] What we are going to show in this section is that selection semantics also faces the paradox of tolerance and that the solution is the same. We can resolve the paradox by thinking about counterfactuals in terms of sufficient similarity rather than precise similarity.

Stalnaker thinks about counterfactuals in terms of **selection models**. These are triples consisting of a set of worlds, a selection set, and a valuation function.

$$\langle W, S, v \rangle$$

The selection set is a set of selection functions, each of which maps propositions and worlds to worlds.[19] Think of these functions as each representing an "opinion" about which worlds are closest.

[17] The proof that rational monotonicity is valid on any transitive frames is basically just the proof from the footnote in section four. That rational monotonicity is valid on *only* transitive frames is also easy to show. Suppose that we have a non-transitive frame. This means that there are worlds $wvuz$ such that $Rwvz$ and $Rvuz$ but not $Rwuz$. Now consider a model with this frame in which $v(A) = \{w, v, u\}$ and $v(B) = \{v, u\}$ and $v(C) = \{v, u\}$. At world z, it is true that $A \boxrightarrow C$ and $A \diamond\!\!\rightarrow B$, but false that $A \wedge B \boxrightarrow C$. So we have our counterexample.

[18] See for example **???**.

[19] More precisely, a selection function is an $f : \mathcal{P}(W) \times W \to W$. The "propositions" here are thus sets of worlds. A propositions is "true at a world" when it has that world as a

$f(A,u) = w$ the closest A world to u is w

Moreover, not only does each selection function have an opinion about worlds are *closest*. They also have an opinion about which worlds are *closer*. That is, each selection function determines a unique closeness ordering on worlds.[20] That ordering is strict in the sense that it has no ties. As far as selection functions are concerned, no two worlds are ever equally close. Finally, that selection models have a *set* of selection functions, rather than just a single selection function, is meant to correspond to the idea that closeness is generally somewhat vague or otherwise underdetermined.

Once we have selection models, the counterfactual operators are defined by supervaluating over the selection functions in the selection set. This means that we need to first define truth relative to a selection function and a world.

$f, w \models A$	iff $w \in v(A)$, when A is atomic
$f, w \models \neg\phi$	iff $f, w \not\models \phi$
$f, w \models \phi \vee \psi$	iff $f, w \models \phi$ or $f, u \models \psi$
$f, w \models \phi \boxrightarrow \psi$	iff $f(\phi, w) \models \psi$
$f, w \models \phi \diamondrightarrow \psi$	iff there is an $f^* \in S$ such that $f^*(\phi, w) \models \psi$

We then say that a sentence is true at a world full stop when it is true at that world relative to all selection functions.[21]

So far so good. We can now establish a certain connection between selection models and similarity relations. In particular, say that a selection function is **compatible** with a similarity relation when:

$f(A,u) = w$ only if $Rwvu$

A selection set S and a similarity relation R then **correspond** when S is the set of all and only the selection functions that are compatible with

member. There are then two further constraints. First, $f(A,w) \in A$. The closest A world is always an A world. Second, if $f(A,w) \in B$ and $f(B,w) \in A$, then $f(A,w) = f(B,w)$. If the closest A world is a B world and the closest B world is an A world, then those worlds are the same.

[20] A world w is closer than v to u according to the selection function f iff there is an $A \subset W$ such that $w, v \in A$ and $f(A,u) = w$

[21] So $w \models \phi$ iff $f, w \models \phi$ for all $f \in S$.

R. What we can show is that each selection set corresponds to a unique similarity relation, and vice versa.

This means is that we can think of selection models as offering a *similarity* analysis of counterfactuals. It's just that facts about similarity are being represented with a set of selection functions, not an explicit similarity relation. The natural question then is: What sort of views about similarity are we committing ourselves to when we use selection models?

The perhaps surprising answer is that we are committing ourselves to the idea that *similarity fails to be transitive*. To see why, consider a simple model. It has four worlds and two selection functions. The selection functions then rank the closeness of worlds to world z as follows:

f_1	f_2
w	v
u	w
v	u
z	z

Worlds that are strictly closer to z are lower on the table. Worlds that are strictly farther from z are higher on the table. Now given a selection set, the corresponding similarity relation is the one such that:

> $Rwvu$ iff there is some $f \in S$ according to which w is closer to v than u

But now looking at the above table, you can see that the corresponding similarity relation is not transitive. There is a selection function according to which w is closer to v and there is a selection function according to which v is closer than u. So $Rwvz$ and $Rvuz$. But there is no selection function according to which w is closer than u. So not $Rwuz$.

Now of course, we can force selection sets to correspond to transitive similarity relations by imposing further requirements. We can require them be **regular** for example.[22] If we do that, transitivity will be regained. Every *regular* selection set corresponds to a transitive similarity relation.

[22]A selection set S is regular iff whenever there is an $f_1 \in S$ according to which w is closer than v and there is an $f_1 \in S$ according to which v is closer than u and $w \neq u$, then

This despite the fact that selection sets in general need not correspond to transitive similarity relations.

As far as I can tell, the standard view is that if we are going to use selection models, then we should assume regularity. We should assume that the selection functions together encode a transitive similarity relation. The question, though, is whether this standard view can be maintained in the face of the tolerance paradox. And the answer, I think, is that it cannot.

Now there is a certain hope that once we make the move to selection models, transitivity can be maintained. The reason is that rational monotonicity is invalid *even if* we require selection sets to be regular and so, in effect, require the corresponding similarity relation to be transitive.[23] This means that our original paradoxical argument is invalid. So it looks like selection models give us a way to maintain the idea that similarity is transitive while also resolving the paradox.

Unfortunately, this hope is misplaced. Regular selection models may invalidate rational monotonicity. But they also validate a different principle called **restricted might transitivity**.[24]

$$\text{RMT} \quad (A \diamondsuit\!\!\rightarrow B) \wedge (A \wedge B \diamondsuit\!\!\rightarrow C) \supset A \diamondsuit\!\!\rightarrow C$$

The problem then is that once we have have this principle, we can run a

there is an $f_1 \in S$ according to which w is closer than u. We are calling this principle regularity because it plays a functionally similar role to a principle that **?** calls regularity.

[23]Consider any model such that $W = \{w, v, u\}$ and $\{f_1, f_2\} \subset S$ and $A = \{w, v, \}$ and $B = \{v\}$ and $C = \{w\}$. Moreover, suppose that $f_1(A, u) = w$ and $f_2(A, u) = v$ and $f_1(A \wedge B, u) = f_2(A \wedge B, u) = v$. In that case, $(A \mathrel{\Box\!\!\rightarrow} C) \wedge (A \diamondsuit\!\!\rightarrow B) \supset (A \wedge B \mathrel{\Box\!\!\rightarrow} C)$ is false at u relative to f_1.

[24]Suppose that $A \diamondsuit\!\!\rightarrow B$ and $A \wedge B \diamondsuit\!\!\rightarrow C$ are both true at world z and fix all similarity and closeness comparisons to that world. So there are closeness relations in S such that the closest A world is a B world and the closest $A \wedge B$ world is a C world. By compatibility, it follows that there is an $A \wedge B$ world w that is at least as similar as any other A world and an $A \wedge B \wedge C$ world v that is at least as similar as any other $A \wedge B$ world. But from this it follows that w and v are equally similar. So by the transitivity of equisimilarity, it follows that v is as similar as any other A world. But this means that there is a closeness relation such that v the most similar A world. Meaning that the there is a closeness relation on which the closest A world is a C world. So $A \diamondsuit\!\!\rightarrow C$ is true at z.

revised version of the paradox without rational monotonicity. First, we reformulate boundedness.

> **Boundedness** ($\diamond\!\!\rightarrow$): There are j and k such that it is false that had Socrates been at least *h*, he might have been at least *k*.

Instead of affirming that Socrates would not have been arbitrarily tall, we simply *deny* that he might have been arbitrarily tall. We then reason to contradiction much like before.

(1)	$\neg(j \diamond\!\!\rightarrow k)$	boundedness ($\diamond\!\!\rightarrow$)
(2)	$j \diamond\!\!\rightarrow j+1$	tolerance
(3)	$n \diamond\!\!\rightarrow n+1$	hypothesis
(4)	$n+1 \diamond\!\!\rightarrow n+2$	tolerance ($\diamond\!\!\rightarrow$)
(5)	$n \wedge (n+1) \diamond\!\!\rightarrow n+2$	4, substitution
(6)	$n \diamond\!\!\rightarrow n+2$	3,5, RMT
(7)	$j \diamond\!\!\rightarrow k$	2,6, induction
(8)	$\bot$	1, 7, PL

Regular selection models, then, may invalidate rational monotonicity, but they do not resolve the paradox.

The solution to the paradox is to deny regularity. This invalidates both rational monotonicity and restrict might transitivity, so invalidates both version of the paradox. But denying regularity means denying transitivity. It means using selection sets that correspond to sufficient similarity relations instead of precise similarity relations. And so the solution for followers of Stalnaker is the same as the solution for followers of Lewis. In both cases, the transitivity of similarity is what has to go.

12. Disconfirmation

Building sensible models on which rational monotonicity can fail is the first step towards resolving the paradox. To have a full resolution, though, we need to explain not just how the inference can fail, but why we mistakenly found it so compelling in the first place.

Something we would like from a theory counterfactuals is a theory of disconfirmation. We like an explanation of how we can—and often

do—convince each other to give up certain counterfactual opinions. For example, suppose I claim that:

> Had Naomi played in the tennis tournament, she would have won.

You, on the other hand, disagree. You say that:

> Had Naomi played in the tennis tournament, she might have lost.

Clearly, there is some sense in which the two of us disagree. The sense in which we disagree, though, will depend on your theoretical commitments. For Lewis, out disagreement is a straightforward disagreement about matters of fact. I believe one thing and you believe something else that is logically inconsistent with what I believe. For Stalnaker, the conflict will be a bit more subtle. Perhaps I take myself to *know* that had Naomi played in the tennis tournament, she would have won. But you think that this is going too far. You think that for all we know—are perhaps for all anyone could ever know—Naomi would have lost.

However we understand the exact nature of our disagreement, what should be clear is that there are certain strategies you might use in order to change my mind. You might, for example, say something like the following:

> Look, I can see where you're coming from. But had Naomi played in the tournament, she might have faced Serena. And had she played in the tournament and played Serena, she would have lost. And so what you say is simply not correct. Had Naomi played in the tournament, she might have lost.

This sort of reasoning would seem to be clearly valid. If you're right that (a) Naomi might have faced Serena and that (b) in that case, she would have lost, then (c) it follows that Naomi might have lost.

The question is, what backs this sort of reasoning? One possibility is that the reasoning is backed by a rule like strengthening.

ST $\quad (A \boxrightarrow C) \supset (A \wedge B \boxrightarrow C)$

That is, whenever you have a true would counterfactual, you can get another true would counterfactual just by conjoining an proposition to the antecedent. In that case, my original claim entails that Had Naomi played in the tournament *and played Serena*, she would have won. But if you're right that Naomi would have lost to if she had played Serena, then this is clearly false. So my original claim must be false too.

The problem is that strengthening is invalid, as we learned from Lewis. If Kangaroos had no tails, the would have toppled over. But if they had advanced prosthetics in place of tails, they would not have toppled over. Had Naomi gone to the party, she would have had a good time. But had Naomi and Sam both gone to the party, Naomi would not have had a good time. But had Naomi and Sam and Serna all gone to the party, Naomi would have had a good time. And so on and so forth.

Lewis's solution to failure of strengthening is to put rational monotonicity in its place. Rational monotonicity is a kind of restricted strengthening principle. It says that we can strengthen from $A \boxrightarrow C$ to $A \wedge B \boxrightarrow C$ under certain special circumstances. Which circumstances are those? The ones in which $A \diamondrightarrow B$. This lets us explain why, for example, it can be true that kangaroos would have toppled over if they had not tails, but would not have toppled over if they had advanced prosthetics. It can be true because it is *false* that if kangaroos had no tails, they might have had advanced prosthetics.

If rational monotonicity were valid, it would explain the validity of your reasoning. After all, I claim that had Naomi played in the tournament, she would have won. You point out that had she played, she might have faced Serena. So it follows by rational monotonicity and my original claim that had Naiomi played in the tournament and faced Serena, she would have won. But we agree that this is false. So on pain of inconsistency, I have to retract my original assertion.

The problem, of course, is that rational monotonicity leads directly to the paradox of counterfactual tolerance. So what we need is an explanation of how you can go about changing my mind that does not depend on rational monotonicity.

What I think is that our counterfactual practice is backed by a rule

you might call disconfirmation.

$$\text{DC} \quad (A \diamond\!\!\rightarrow B) \wedge (A \wedge B \mathrel{\Box\!\!\rightarrow} C) \supset (A \diamond\!\!\rightarrow C)$$

This rule directly validates your earlier reasoning. Had Naomi entered the tournament, she might have faced Serena. Had she entered the tournament *and* faced Serena, she would have lost. So it follows that had Naomi entered the tournament, she might have lost. This precisely because she might have faced Serena.

The advantage of disconfirmation is that it lets us steer clear of the paradox. We know this because disconfirmation is valid over the class of all near similarity models.[25] It is also valid over the class of all selection models.[26] But since rational monotonicity is invalid on both classes, it follows that we can accept disconfirmation while denying rational monotonicity.

Moreover, you can see where falsification might easily be *confused* with rational monotonicity, and so explain why we found rational monotonicity so compelling in the first place. Given duality *and* the counterfactual law of the excluded middle, disconfirmation and rational monotonicity are logically equivalent. For followers of Lewis, this means

[25] Suppose $A \wedge B \mathrel{\Box\!\!\rightarrow} C$ at z and fix all near similarity comparisons to z. This means that there is an $A \wedge B$ world w such that all the $A \wedge B$ worlds that are at least nearly as similar are C worlds. Now suppose for reductio that $A \mathrel{\Box\!\!\rightarrow} \neg C$. There is thus an A world v such that all of the A worlds that are at least nearly as similar are $\neg C$ worlds. So in particular, w is not at least nearly as similar as v. Next, suppose that $A \diamond\!\!\rightarrow B$ at z. From this it follows that there is an $A \wedge B$ world u that is at least nearly as similar as v. But then since all of the the $A \wedge B$ worlds that are at least nearly as similar as v are $\neg C$ worlds, it follows that u is a $\neg C$ world. This means that u is not at least nearly as similar as w. But we already said that w is not at least nearly as similar as v. So by acyclicality, it follows that u is not at least nearly as similar as v. But this is contrary to assumption. So $\neg(A \mathrel{\Box\!\!\rightarrow} \neg C)$ at z and therefore $A \diamond\!\!\rightarrow C$ at z.

[26] Suppose that there is some $f \in S$ according to which the closest A worlds is a B world. Suppose, moreover, that the closes $A \wedge B$ world is a C world according to every $g \in S$. It then follows that the closest $A \wedge B$ world is a C world according to f. But given the constraints governing selection functions, this means that the closest A world according to f is identical to the closest $A \wedge B$ world. So there is some $f \in S$ such that the closest A world is a C world.

that whatever explains our (mistaken) attraction to the counterfactual law of the excluded middle can also explain our (mistaken) attraction to rational monotonicity. For followers of Stalnaker, something similar is true. Whatever explains our (mistaken) attraction to duality can also explain our (mistaken) attraction to rational monotonicity. In both cases, we can just reduce the one mistake to another mistake we are already committed to explaining.

13. Pollock

John Pollock suggests that we should analyze counterfactuals using what we are going to call **Pollock models**.[27] His models invalidate rational monotonicity, so would seem to offer a way to resolve the tolerance paradox. In the section, I want to say why I think that sufficient similarity models are preferable.

Like sufficient similarity models, Pollock models are triples consisting of a set of worlds, a counterfactual accessibility relation, and a valuation functions.

$\langle W, R, v \rangle$

There are then two key differences. The first is that Pollock does not require his accessibility relation to be total and acyclic. Instead, he requires it to be reflexive and transitive.

Reflexive	$Rwwz$
Transitive	if $Rwvz$ and $Rvuz$, then $Rwuz$

The second difference is that Pollock defines the counterfactuals somewhat differently. When we constructed sufficient similarity models, we followed Lewis and used the first of the below definitions. Pollock uses the second. The difference when it comes to might counterfactuals is similar.

[27] See for example **????**.

$u \models A \boxdot\!\!\rightarrow B$ iff there is a world $v \in W$ such that $v \models A$ and every world $w \in W$ is such that $w \models A \supset B$ whenever $Rwvu$

$u \models A \boxdot\!\!\rightarrow B$ iff there is a world $v \in W$ such that $v \models A$ and every world $w \in W$ is such that if $w \models A \wedge \neg B$, then not both $Rwvu$ and not $Rvwu$

As it turns out, these two difference cancel out and the result is a kind of formal equivalence. Pollock models and sufficient similarity models validate all the same rules and axioms. So from a purely formal perspective, I have no beef with Pollock.

The problem is making sense of his accessibility relation. Pollock thinks of his accessibility relation the **containment of change** relation.

$Rwvu$ the changes needed to get from u to w are a subset of the changes needed to get from u to v

What this means that when $Rwvu$ and not $Rvwu$, the changes needed to get from u to w are a *strict superset* of the changes needed to get from u to v. To get from world u to world w, you have to do everything you need to do to get from u to v, and then some.

For my own part, I am not sure that I understand the containment of change relation. And insofar as I do, I strongly doubt that it has the right extension. For example, the containment of change relation would seem to stand in the way of solving the tolerance paradox. Here is why: Suppose that Socrates is in fact five feet tall. Now consider any pair of worlds w_n and w_{n+1} that are otherwise similar, but at which Socrates is n and $n+1$ Planck lengths taller, respectively. Insofar as I understand the containment of change relation, the changes need to get from the actual world to w_{n+1} are a strict superset of the changes need to get from the actual world to w_n. To make Socrates $n+1$ Planck lengths taller, you have to do everything you have to do to make him n Planck lengths taller, and then some, with the same going for every other such pair of worlds. But in that case, tolerance fails. Had Socrates been at least six feet tall, he

would not have been even a Planck length taller.[28]

What this shows, it seems to me, is that we should scrap Pollock's containment of change relation and replace it with something else. But what? One idea would be to go back to using precise similarity. We could say that:

$Rwvu$ w is at least as similar as v to u

Failures of totality would then correspond to failures of comensurability. Sometimes, there are pairs of worlds w and v such that w is neither strictly more similar, nor strictly less similar, nor equally similar.

The problem is that failures of rational monotonicity will then require failures of commensurability. So for example, to block the tolerance paradox, we will need pairs of worlds w_n and w_{n+1} whose similarity to the actual world is incommensurable. But why would they be *incommensurable*? You would think that w_n would be strictly more similar than w_{n+1}. Or perhaps they should count as equally similar because a single Planck length is too small to make a difference. Either way, though, their similarity is commensurable. What we need is some reason to think that their similarity is *incommensurable*. And I have no idea what that could be.

Maybe if we fumbled around long enough, we could save Pollock models. We could find a relation with both the right formal properties and the right extension. A simpler solution, though, is to switch to sufficient similarity models. Sufficient similarity models validate all the same rules and axioms. But in the case of sufficient similarity models, we already have a relation that would seem to have both the the right formal properties and the right extension. So why bother?

[28] A similar point applies to **?**. She suggests a variation of Pollock's semantics on which the containment of change relation is fixed using what she calls an ordering source. But while her semantics validates the same rules and axioms as sufficient similarity models, the problem is finding a natural order source that gives us both tolerance and boundedness. For my own part, at least, I have been unable to find one.

14. Independence

Pollock's own reasons for denying rational monotonicity are not especially compelling. The cases in his (**?**) and (**?**) are somewhat baroque, so we can instead consider a simper version from **?**.

Suppose we have three characters Alice, Bernie, and Carole who flip three fair coins. Those flips are causally independent. As it happens, Alice and Carole both flip heads. Bernie flips tails. We then consider three counterfactuals:

(1) Had Alice and Bernie flipped the same, Carole would have flipped heads.

(2) Had Alice and Bernie flipped the same, Alice and Bernie and Carole might have all flipped the same.

(3) Had Alice and Bernie and Carole flipped the same, they would have all flipped heads.

The first two are true, but the third is false, or so you might think. Why is that? Well, suppose you accept a certain plausible principle connecting counterfactuals and causation.

> **Independence:** If $\neg A$ and B are both true at the actual world, then $A \mathrel{\square\!\!\rightarrow} B$ iff those facts are causally independent.

This principle then entails that the first two counterfactuals are true and that the third is false. So rational monotonicity is invalid.

Here is how the reasoning goes: The first counterfactual is true because changing how Alice and Bernie flipped would not have changed how Carole flipped, and she in fact flipped heads. The second is true because had Alice and Bert flipped the same, that might have been because Bert flipped heads. And by independence again, Carole would still have flipped heads. So they all might have flipped heads and so might have all flipped the same. But the third counterfactual is false because all three flipping the same is not causally independent from all three flipping heads. All three flipping heads *just is* a way of all three flipping the same.

The problem for Pollock is that independence-style reasoning not only predicts the invalidity of rational monotonicity. It also predicts the

invalidity of other principles like **restricted strengthening** that are validated by his semantics.[29] So if we are going to deny rational monotonicity on the basis of independence, Pollock's semantics is not the way to do it. We are going to need a much more radical departure from Lewisian semantics.[30,31]

The advantage of the tolerance paradox is that we can motivate the denial of rational monotonicity without appealing to anything like causal independence. And so we can just sidestep the whole problem.

15. Dynamic Semantics

One especially simple view is that would counterfactuals are strict conditionals. $A \boxrightarrow B$ is equivalent to $\Box(A \supset B)$. Might counterfactual are then understood as expressing certain kinds of possibilities. In particular, $A \Diamond\!\!\rightarrow B$ is equivalent to $\Diamond(A \wedge B)$. Like Lewisian semantics, the strict conditional view validates all the rules used in the tolerance paradox.[32] This means that defenders of the strict conditional view will have to either deny that counterfactuals are tolerant or deny that they are bounded.

That may be. But nowadays, defenders of the strict conditional often accept certain dynamic mechanisms for shifting contexts. To say that $\Box(A \supset B)$ is to say that $A \supset B$ is true at all the relevant worlds. Since the relevant worlds can shift from context to context, the truth of counterfactuals can shift from context to context. This gives defends of the strict conditional analysis additional powerful tools for explaining away counterexamples. You might think, then, that these sorts of tolls might help with the tolerance paradox.

[29] Restricted strengthening says that $(A \boxrightarrow B) \wedge (A \boxrightarrow C) \supset (A \wedge B \boxrightarrow C)$. The observation that independence entails the failure of restricted strengthening is from **?**.

[30] I owe this point to Simon Goldstein.

[31] You could think that the relevant judgements are correct, but deny that this on the basis independence. Myself, I have a hard time hearing the first two sentences as true and the third as false. And when I can get those readings, I strongly suspect that I'm simply smuggling in independence-style reasoning.

[32] More precisely, it validates substitution, duality, and rational monotonicity so long as $\Box$ is a normal modal operator and $\Box\phi$ is equivalent to $\neg\Diamond\neg\phi$.

To see how context shifts can help, consider apparent counterexamples to strengthening, an especially potent cousin of rational monotonicity that we met in section 12. Where rational monotonicity says that we can strengthen from $A \boxrightarrow C$ to $A \wedge B \boxrightarrow C$ under certain special circumstances, strengthening says that we can strengthen under any circumstances whatsoever. Strengthening is validates by the strict conditional analysis, but would seem to have clear counterexamples. To use the case from Lewis again, if kangaroos had no tails, they would have toppled over. But if kangaroos had advanced prosthetics in place of tails, they would not have toppled over.

One strategy for explaining away such counterexamples is suggested by **?**. Here is the basic idea: Suppose we start in a context in which the only relevant possible worlds in which kangaroos have no tails are worlds in which they topple over. Furthermore, there are no relevant worlds in which they have advanced prosthetics. You then assert that if kangaroos had no tails, the would have toppled over. This is true, given the the strict conditional analysis. Now you go on to assert that if kangaroos had advanced prosthetics in place of tails, the would not have fallen over. This assertion, von Fintel thinks, presupposes that the context includes worlds in which kangaroos have advanced prosthetics. Since the original context has no such world, the context shifts to accommodate. In particular, it shifts by adding all of the most similar worlds in which kangaroos have advanced prosthetics to the set of contextually relevant worlds. Since those are worlds in which kangaroos do not topple over, your second assertion is true.

So far so good. The question now is whether a similar mechanism can explain away the tolerance paradox. To fix on an example, suppose that we start out in a context that includes worlds in which Socrates is up to seven feet tall. I then assert that had Socrates been at least six feet tall, he would have been no taller than seven feet. This true, given the context, so we get boundedness. The question is whether we can explain tolerance. Suppose I claim that had Socrates been at least seven feet tall, he might have been a Planck length taller. This is obviously true, but is false on the strict conditional analysis, since the context does not include any worlds

in which Socrates is taller than seven feet.

One idea, suggested by **?**, is that asserting a might counterfactual $A \diamond\!\!\rightarrow B$ presuppose that the context includes worlds at which $A \wedge B$. If the context has no such worlds, it accommodates the presupposition by adding the most similar $A \wedge B$ worlds. This is how that might help with the previous example: I assert that had Socrates been at least seven feet, he might have been at least a Planck length taller. There are no worlds where Socrates is a Planck length taller than seven feet in the original context, so the context shift to accommodate by adding the most similar such worlds. My assertion is then true in the new context. So it would seem, we have a dynamic strategy for explaining away the tolerance paradox. Counterfactuals seem tolerant because contexts always accommodate might counterfactuals.

A serious problem with the proposed mechanism is that it is overly accommodating. If I say that had Socrates been at least seven feet, he might have been at least eight feet, the context will shift to accommodate. Sure. That's fine. But for the same reason, if I say that had Socrates been at least seven feet, he might have been a million feet, the context will *also* shift to accommodate. But this is the wrong result. Had Socrates been at least seven feet, he would not have been a million feet. Or take another example. Suppose you make the following speech:

> Had Socrates been at least a million feet, he might have been a million feet. But had he been at least six feet, he would have been less than ten feet.

The first assertion presuppose that there is a contextually relevant world in which Socrates is a million feet tall. So if the context has no such world, it accommodates by adding one. But in that case, your second assertion is false. But again, this is the wrong result. Saying that had Socrates been a million feet, he would have been a million feet should have no bearing on whether you can go on to say that had Socrates been at least six feet, he would have been less than ten feet.

Perhaps we should not read too much into such difficulties. These are only objections to a particular mechanism for shifting contexts, and there may well be more sophisticated mechanisms that can do the job.

Still, to actually solve the tolerance paradox, we would need to specify those mechanisms. And it is surprisingly hard to see what they might be.[33]

16. Conclusion

The first half of this paper presented the paradox of counterfactual tolerance and the second presented my own preferred solution. I think that what the paradox shows is that we need to analyze counterfactuals using a non-transitive similarity relation. We need to use sufficient similarity in place of precise similarity.

All of that said, I am more committed to the *paradox* than am to any particular solution. What I think is that any plausible theory of counterfactuals will have to resolve the paradox one way or another and that—however the paradox is resolved—we will have to say something interesting. The paradox thus serves as an important constraint on what our final theory of counterfactuals can look like.

[33] Similar skepticism about dynamic semantics is raised in **?**.