Nearby worlds

Wolfgang Schwarz <wo@umsu.de>

Draft, February 28, 2007

Abstract. I discuss several classes of counterexamples to the Stalnaker-Lewis analysis of counterfactuals, and attempt to defend it by appealing to general semantic and pragmatic considerations.

1 Counterfactuals and similarity

If I drilled a hole through the wall behind my desk, I would end up in Bob Goodin's office. That's because Bob Goodin is in the office next door, and my desk faces this wall. If things were arranged otherwise, the counterfactual might well be false . In this respect, counterfactual implication differs from strict implication: when we evaluate the counterfactual in the present situation, we consider what happens at worlds where I drill a hole through the wall behind my desk, but we don't consider *all* such worlds. Roughly speaking, when we evaluate the counterfactual in a given situation S, we ignore possibilities that differ gratuitously from how things are in S over and above the difference that comes with the truth of the antecedent.

Thus we reach the Stalnaker-Lewis analysis of counterfactuals:

> $A \mathrel{\square\!\!\rightarrow} C$ is true in S iff C is true at the A-worlds otherwise most similar to S.[1]

Now, it is well-known that intuitive judgements of overall similarity often deliver the wrong results when plugged into this analysis (see Fine 1975 and many others):

> **1)** if Nixon had pressed the button, there would have been a nuclear holocaust.

This is true despite the fact that worlds where the button is pressed and the mechanism fails are in many ways more similiar to the actual world than worlds where the button is pressed and the mechanism works.

The example shows that certain respects of similarity can easily trump others: in evaluating (1), we hold fixed that the mechanism works; we ignore worlds where it doesn't, no matter how similar they are otherwise to the actual world.

So what are the relevant standards of similarity? Here is Lewis' proposal (1979, pp.47f., see also his 1973, pp.75–77):

1 I presuppose the 'Limit Assumption' here, but not the 'Uniqueness Assumption'. Reasons for the latter will appear soon; I don't have any reasons for the former, but also I don't understand Lewis's reasons for rejecting it (see his 1981: 89), and the assumption simplifies things here.

(1) It is of the first importance to avoid big, widespread, diverse violations of law.

(2) It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.

(3) It is of the third importance to avoid even small, localized, simple violations of law.

(4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

A world where Nixon presses the button and the mechanism fails, Lewis explains, is approximately similar to the actual world for a longer stretch of time, but according to clause (4), this doesn't count for much, not enough to outweigh the "small miracle" of the mechanism failing. Indeed, this seems to be one of the cases where approximate similarity has *zero* weight: assume there was a loose contact in the button mechanism, with a 0.5 chance of failure. In this case, (1) is intuitively false, but so is

2) if Nixon had pressed the button, there would not have been a nuclear holocaust.

But this time, the increase in approximate similarity is not even compensated a small miracle. So approximate similarity should count for nothing here. (The example is similar to an example by Tichy 1976, which Lewis also handles by giving zero weight to the similarity, see 1986, p.65, fn.10.)

In other cases, approximate similarity seems to matter, and several philosophers have come up with promising explanations of what distinguishes zero-weight similarity from similarity that counts (see Johnson 1991, Bennett 2003, ch.15, Edgington 2003, Schaffer 2004, Kment 2005; Lewis himself only notes that "[d]ifferent cases come out differently").

I don't want to get into these details here. I mention all this only to emphasize that "similar" in the Stalnaker-Lewis analysis is a technical term that, once fleshed out, may have little or nothing to do with our intuitive concept of overall similarity.

2 *A*-worlds

So $A \boxdot\!\!\rightarrow C$ is true iff C is true at the selected A-worlds, where 'selection' at least partly depends on the world of evaluation, perhaps in the way Lewis suggested (or perhaps in some other way).

What is an A-world? What is a world where A (or C) is true? It is not, presumably, a world where the *string of symbols*, irrespective of its meaning, is true. Worlds where water is wet and "light is slower than sound" means that water is wet do not count as worlds where "light is slower than sound" is true. One is tempted to say that an A-world is a world where *the proposition actually expressed* by A is true. However, this may just lead us around in a circle: the relevant 'proposition' here is the sentence's *secondary*, or *horizontal*, or *C-proposition*, standardly defined as something like the set of worlds – 'considered as counterfactual' – at which the sentence is true.

But let's set this question aside.[2] What's important is that the evaluation of sentences at counterfactual worlds is different in many ways from the evaluation of sentences at utterance contexts. This is obvious for sentences involving indexicals, but also for ambiguous sentences:

2 I think C-propositions should be derived from A-intensions by means of compositional rules, telling us, e.g. that

3) If Fred had been to the bank yesterday, he could pay us back now.

What if Fred actually was at the river bank yesterday, but nowhere near a money bank? Does the actual world then qualify as a selected A-world? No. But not because it is less 'similar' to itself than other worlds; rather because in the context of (3), worlds where Frad wasn't at a money institute don't count as A-worlds at all. What matters is not that the English sentence "Fred has been to the bank" is true at a world, but rather that the disambiguated (secondary) proposition expressed by it in the utterance context is true there.

Unfortunately, things are more complicated. The antecedent doesn't always select worlds satisfying its secondary intension: "if the IAU had decided that Pluto is a planet" — here we consider worlds where "planet" means something else; "if water had turned out to be XYZ", "if the morning star was different from the evening star" — here we appear to consider worlds satisfying primary propositions; still stranger things happen in "if I were you", "if I found a proof that ZFC is inconsistent" and "if 12 was prime". In what follows I will assume that the selected A-worlds are at least worlds where A is true; but it is worth keeping in mind that even this isn't universally true.

What is true for ambiguities arguably also holds for other kinds of indeterminacy.

4) The [rugby] ball would be easier to catch if it was round.

5) If we had left early, we would have reached the summit before the storm.

Context determines what counts as "round" and "early" here; in (4), we don't consider worlds where the ball is perfectly round. In (5), we don't consider worlds where we left only a minute earlier, nor worlds where we left last week.

In both cases, we would get the wrong result if we let the indeterminacy be resolved by similarity: worlds where we leave only a minute earlier are presumably at least as close as worlds where we leave about an hour earlier. The former worlds are ignored, it seems, not because they are more remote, but because they aren't A-worlds in the first place; they aren't worlds where the proposition expressed by the antecedent in the context of (5) is true. (It wouldn't help to appeal to coarse-grained similarity standards on which 1-minute-earlier worlds are equally close to actuality than 1-hour-earlier worlds: that would still render (5) false.)

Likewise, in (4), we only consider worlds where the ball is significantly rounder than it actually is, even if in other contexts, something with its current shape would already qualify as "round".

Vague terms usually remain vague even in a particular utterance context. The contexts of (4) and (5) rule out certain shapes and times as round and early, but they don't draw a precise line. If in a given context, a sentence is still indeterminate, we usually regard it as true, or assertible, only if it is true on all resolutions of the (remaining) indeterminacy. Thus if a counterfactual has a vague antecedent, we regard it as assertible only if it is true on all resolutions of the

---

w is in the C-proposition of "a is F" iff the A-extension of a, let's call it x, and the A-extension of F, let's call it y, are such that w represents x as having y. How do we know that we have found the correct rules? By testing the result against our judgements about (and, more generally, against the A-intensions of) counterfactuals and other modal sentences; not by testing them against off-hand judgements about 'the proposition expressed', 'what is said', etc., as it is doubtful whether propositions suiting *that* job description will do for the semantics of modal embeddings.

indeterminacy, no matter if some resolutions require us to consider more remote worlds than others.

> **6)** # If there was a heap of sand in the backyard, it would consist of 7385 grains of sand.

This pattern turns out to be very widespread.

3 Antecedent spreading

In an ordinary context, (6) is false because, intuitively speaking, the antecedent can be true in many ways, but the consequent is true in only one of those.

This is an instance of a phenomenon I call *antecedent spreading*, which always occurs when the antecedent of a conditional remains explicitly neutral between several more specific propositions. Classic examples involve disjunction:

> **7)** If I had arrived 5 or 10 minutes later, I would still have caught the train.

We take this to entail that if I had arrived 10 minutes later, I would still have cought the train. But

$$\frac{A \vee B \mathrel{\Box\!\!\rightarrow} C}{B \mathrel{\Box\!\!\rightarrow} C}$$

is invalid in the Stalnaker-Lewis framework. If C is true at the closest A-world but false at the closest B-world, and the closest A-world is closer to actuality than the closest B-world, $A \vee B \mathrel{\Box\!\!\rightarrow} C$ is true and $B \mathrel{\Box\!\!\rightarrow} C$ is false. (Unless the selection relation is antecedent-dependent, the rule entails [and is equivalent to?] the rule of strengthening the antecedent.)

This is the first class of apparent counterexamples I will discuss.

More examples:

> **8)** If Fred had drunk beer or wine that night, he would have died.

> **9)** if Fred had drunk from a bottle on the top shelf, he would have died.

We take this to entail that for any bottle on the top shelf, if Fred had drunk from it, he would have died.

> **10)** If the dart had landed on a red field or anywhere on the left-hand side of the board, I would have won.

Again, we take this to imply not only that I would have won in the closest among the mentioned possibilities, but in all of them.

A variation of a famous example from Goodman (xxx):

> **11)** If Jones was dropped in Korea today, he would soon be fighting in the war.

Background: Korea is at war and divided into North Korea and South Korea. In this setting, (11) is acceptable only if Jones would be fighting no matter which part of Korea he was dropped at. Otherwise one should reply: "it depends on whether he is dropped in NK or SK".

> **12)** If I had been run over by a car yesterday, I would be either dead now or severely injured.

This implies that on any (reasonable) way I could have been run over by a car, I would now be dead or severely injured. To see the implication, drop one of the disjuncts in the consequent and notice how it sounds false. Likewise in:

> **13)** If Vladimir Putin had played Russian roulette on June 6, 2005, he would have either survived or died.

The fact that a world where Putin survives is more similar to the actual world than one where he dies is irrelevant: the counterfactual asks us to consider both outcomes. Lewis could also handle this case by discounting approximate similarity, but this wouldn't help in the following case (from Stephan Leuenberger):

> **14)** # If lots of big miracles had occurred yesterday, everything would be just as it actually is today.

Big miracles can remove all the traces of earlier miracles. So there are worlds where lots of big miracles occurred yesterday, yet from today on everything is exactly as it actually is. Following Lewis's similarity standards, one would expect these worlds to be closer to actuality than worlds where the future is affected by the miracles; so (14) would be true. But it is not. And intuitively, the reason why it is not is that there are many ways the antecedent could be true that would make the consequent false.

Antecedent spreading is a special case of what one might call *diamond spreading*. A 'diamond context' is a context that in one way or another involves existential quantification over possibilities. Consider the epistemic "might": "it might be that *A*" normally means that *some* epistemically possible worlds are *A*-worlds. But if *A* leaves certain facts explicitly open, we take it that all the options are epistemically possible.

> **15)** Fred might have had beer or wine

implies that Fred might have had beer and Fred might have had wine.

> **16)** Fred might have tried one of bottles on the top shelf

implies that he might have tried any of the bottles.

Or take permission: "*A* is permissible" normally means that some permissible worlds are *A*-worlds. But again,

> **17)** You may have beer or wine

implies that you may have beer and that you may have wine. And

> **18)** You may try one of the bottles on the top shelf

normally implies that we may try any one of the bottles, not just a particular one or two of them.

Or take nomological possibility:

> **19)** It is nomologically possible for this cup to move sideways or even upwards when dropped.

implies that it's nomologically possible for the cup to move upwards when dropped.

Not surprisingly, this phenomenon has attracted some attention in formal semantics, but I'm not aware of any really satisfactory account. Some proposals only deal with disjunction and thus aren't general enough; on the other hands, some proposals can't explain why the spreading doesn't happen in box contexts: "it is necessary/obligatory that A or B" doesn't imply that it is necessary/obligatory that A. (This raises the question why the antecedent of conditionals behaves like a diamond context. Perhaps because it is a negated compound in a box sentence?)

[**?**] points out an interesting fact: suppose the inference from $A \vee B \boxrightarrow C$ to $B \boxrightarrow C$ is generally valid; then provided we may substitute truth-functional equivalents in the antecedent of a counterfactual, strengthening the antecedent – from $A \boxrightarrow C$ to $A \wedge B \boxrightarrow C$ – would also have to be valid, for $A \boxrightarrow C$ is equivalent to $(A \wedge B) \vee (A \wedge \neg B) \boxrightarrow C$. But strengthening the antecedent does not usually seem valid. Fine suggests that the "or" in disjunctive antecedents may be a special conjunctive "or" with which A is not equivalent to "$(A \wedge B)or(A \wedge \neg B)$". Fine also points out that the problem carries over to permission and preference, and makes some further proposals.

In other cases, the effects of valid diamond spreading would be even worse: if "permissible(A)" entails "permissible(AvB)" and "permissible(AvB)" entails "permissible(B)", either nothing or everything would have to be permissible. The fact that in standard deontic logic P(A) entails P(AvB) is sometimes calles "Ross's Paradox".

Here is my attempt at a diagnosis. Let's call the relevant predicates *range predicates*. If we predicate a range predicate of several things, then we often implicate that the range is distributed over those things:

> **20)** The rooms are between 12.7 and 16.5 square meters.

> **21)** The best things in life are illegal, immoral or fattening.

(The implicature is also present in "all..." and "every...", but not as strong. To inform you that the smallest room is 12.7 and the largest 16.5, "the rooms are between..." is preferable to "all rooms are between..." and "each room is between...".)

Not surprisingly, this is still so if the relevant things are introduced by a (plural) quantifier:

> **22)** I looked at some rooms of between 12.7 and 16.5 square meters.

> **23)** Some of the guests gave only 1 or 2 dollars.

(Many "some Fs are A or B" sentences sound strange, like "some (of the) guests gave 1 or 100 dollars". That's because the relevant plurality that would make this true is gerrymandered. "Some Fs are between x and y" usually sounds better.)

24) Sometimes she brought mushrooms or flowers.

"Sometimes" is a plural quantifier; the example doesn't work with "at (at least) one time". It does work with "a few times", "at most times", "often", "always". But maybe these cases aren't all alike: in the scope of "sometimes" and "at many times", "often", you can refer to the selected times or operate on them in the matrix: "sometimes, she is sad and *then* (at those times), she always brings coffee". You cannot do so in "always". Maybe "always" is like "all", where we also have a distributing effect, but a weaker one and for different reasons.

(Many assertions that don't distribute are more natural with "always/all" than with the other locutions: "it does always either rain or not rain" – "at some times, it either rains or not".)

(Is "all" a plural quantifier? Sometimes it seems to be one: "all humans have a common ancestor". (Compare the nonsensical "each" counterpart.))

The fact that we can backrefer to the relevant times also shows that (24) cannot be analysed as

25) Sometimes she brought mushrooms and sometimes she brought flowers.

For there is no plausible analysis along these lines for

26) sometimes, the sun was shining and then she always brought mushrooms or flowers.

This doesn't mean the same as

27) sometimes, the sun was shining and then she always brought mushrooms, and sometimes, the sun was shining and then she always brought flowers,

or

28) sometimes, the sun was shining and she brought mushrooms, and sometimes, the sun was shining and she brought flowers.

The next step is now pretty obvious: since "may", "might", "possibly" have a lot in common with "some", they share this pragmatic feature: saying that Fred might be at the pub is saying that at *some* (doxastically) possible worlds, he is at the pub. Thus we get the or-to-and inferences for sentences like

29) Fred might be in the library or at the pub;

30) you may have beer or wine.

31) you may have a drink from the top shelf.

The 'range predicate' is here a range proposition that is 'predicated' of a collection of worlds. The effect is that we interpret the predicate to distribute over this collection, especially if we assume the speaker to have some knowledge about what is and what isn't part of the collection.

Again, there are examples where we backrefer or operate on the selected collection which a conjunctive analysis can't handle:

32) Fred might have left already, and then he will certainly be either in the pub or in the library.

(I'm assuming that "Fred might be at the pub" is not to be analysed with the singular "in *at least one* world, Fred is at the pub". Is there other evidence for this?)

Why does the implicature not arise for box contexts such as

33) I know that Fred is either in the pub or in the library,

and

34) you must hand in the paper within 2 weeks,

which don't entail that I know that Fred is in the pub and that you must hand in the paper after exactly 14 days? The answer is that it does. Here, too, it is implicated that the range proposition distributes over the selected cases – the doxastically possible worlds or the legally acceptable worlds. Thus (33) implicates that at some of my doxastically possible worlds, Fred is in the pub, and (34) implicates that you may hand in the paper after exactly 14 days.

If one thinks of the phenomenon as an or-to-and inference, it looks like it only affects diamond contexts, but if one thinks of it as a distribution-over-all-selected-items implicature, it applies throughout.

Finally, why does the implicature show up in the antecedent of counterfactuals? It also shows up in relative clauses restricting universal quantifiers, which are very similar constructions (the counterfactual can be analysed as "in all the closest worlds where..."):

35) all rooms between 12 and 16 square meters are available for rent

implies that some room with 12 square meters is available for rent. I don't really know why. Meybe it's because "all F are G" implies "some things are F and G", to which the previous explanation would apply.

If diamond spreading is a matter of implicature, it should be cancellable. And indeed, in some cases you can enforce the semantically expected interpretation by saying a bit more:

36) It is nomologically possible for the plate to go sideways, or even upwards – I can't remember which.

37) You may have beer or wine – which of them I will tell you later.

38) if Fred had drunk from a bottle on the top shelf – I think it was the Martini bottle, but I'm not sure –, he would have died.

But the last case is already strained, and in many of the above examples, I can't think of a convincing cancellation. Moreover, these apparent cancellations might also just be disambiguations. After all, you *can* say

39) if Fred had drunk from a bottle on the top shelf, it would have been the Martini.

(Then again, all such cases sound a bit strange, like the 'backtracking' counterfactuals Lewis discusses in 1070: they don't really explore the consequences of a counterfactual assumption.)

So I don't have a good explanation of diamond spreading. But even without an explanation, the phenomenon plainly exists. The upshot is that if the antecedent A of a conditional $A \boxdot\!\!\rightarrow C$ explicitly leaves it open whether B_1 or B_2 or ..., then we normally regard the conditional as true iff all the $A \wedge B_i \boxdot\!\!\rightarrow C$ are true. We're considering *more* possible worlds than expected: we're not only considering the closest worlds where the antecedent is true in one way or another; instead, the range of selected worlds extends to comprise at least one world from all the ways the antecedent could be true.

It is certainly not true that whenever the antecedent of a conditional leaves something open, we have to consider all ways of filling in what is left open. "If I had looked at the door" doesn't say anything about the arrangements of things in the room or about a tendency of my computer screen to always remain in sight, but we don't evaluate the consequent and worlds that differ from actuality in these respects.

That's why I said the antecedent must leave the relevant questions *explicitly* open, in the way a disjunction and an existential quantification does. This is of course extremely rough, but I suspect a better account will have to wait until we have a general explanation of diamond spreading.[3]

The present case is subtly different from cases involving vague terms where we treat the counterfactual as determinately true iff it is true on all resolutions of the vagueness. First, in no ordinary sense is a disjunction $A \vee B$ *indeterminate* or *ambiguous* between its disjuncts A and B, and it would be ridiculous to apply supervaluationism here, claiming that $A \vee B$ is true iff it is true on both resolutions of the 'indeterminacy' – making it equivalent to $A \wedge B$.

Moreover, vagueness and antecedent spreading can combine:

44) # if there was a heap of sand in the backyard, it would be a very small heap.

This sounds false, but it may well be that on all resolutions of the vagueness of "heap", the closest world where there is a heap in the backyard is one that falls under the corresponding

---

3 A less obvious, but perhaps related phenomenon is that what is *presupposed* by a consequent is less in doubt than what is openly *stated*:

40) If I had bet on heads, I would have won.

41) ? If I had bet on heads, the coin would (still) have landed heads.

42) If somebody came down the street now, they would think we're crazy.

43) ? If somebody came down the street now, they would not be blind.

With the assumption left implicit, we resolve/enrich the antecedent in such a way that the counterfactual makes sense, viz. by holding fixed the outcome of the coin toss and assuming that the 'somebody' is some normal person with working eye-sight. But we can't do that any more if the assumption is turned into an explicit consequent: it isn't charitable to interpret the utterer of (43) as claiming that if somebody with working eye-sight came down the street now, they would not be blind. This claim would be too trivial to utter. So we widen the class of antecedent cases, making it doubtful whether the counterfactual is true.

resolution of "very small heap". (44) asks us to consider all reasonable ways of how there could be heap of sand in the backyard, not just the closest such ways.

The difference also shows up when we turn from 'would'-counterfactuals to 'might', 'could' and 'probably'-counterfactuals:

> **45)** If Vladimir Putin had played Russian roulette, he could have died.

> **46)** If lots of big miracles had occurred yesterday, things would probably be different now.

If "Vladimir Putin played Russian roulette" was somehow indetermined between "... and died"/"... and survived", and we're considering both resolutions of the indeterminacy when evaluating (13), we should also consider both resolutions when evaluating (45). But then (45) would clearly be false, which it is not. So it seems that we really have a single set of antecedent worlds here, not indeterminacy between different sets.

Here I assume – what is controversial – that might-counterfactuals can be interpreted as dual of would-counterfactuals:

> "if kangaroos had no tails, they *would* topple over" is true iff they topple over in *all* closest A-worlds;
>
> "if kangaroos had no tails, they *might* topple over" is true iff they topple over in *some* closest A-world;
>
> "if kangaroos had no tails, they *would probably* topple over" is true iff they topple over in *most* of the closest A-worlds;
>
> etc.

The counterfactual antecedent is here in effect regarded as a plural descriptor – at *the closest A-words*, C – rather than as a singular descriptor – at *the closest A-world*, C, (see Lewis, *Counterfactuals*, xxx, Schlenker, "Conditionals as Definite Descriptions") and the other locutions are other generalized quantifiers.

It seems that there are other uses of might- and probably-counterfactuals, on which "if A, M C" expresses epistemic possibility of $A \mathrel{\Box\!\!\rightarrow} C$ (see Stalnaker 1984) and on which it expresses that $A \mathrel{\Box\!\!\rightarrow}$ it is nomologically possible that C (see Lewis 1986). (Example: suppose the laws of nature either entail $A \rightarrow C$ or $A \rightarrow \neg C$, but it is unknown which. Then "if A, it might be that C" and "if A, it might be that $\neg C$" appear true, but so is "either $A \mathrel{\Box\!\!\rightarrow} C$ or $A \mathrel{\Box\!\!\rightarrow} \neg C$", contradicting duality.)

4 Antecedent strengthening

I've suggested that in a suitable context, the antecedent of

> 5) if we had left early, we would have reached the summit before the storm.

expresses a proposition that is only satisfied in worlds where we left *at least an hour* earlier than we actually did. But this story doesn't easily carry over to many related cases.

> **47)** If Fred was taller, he could reach the book on the top shelf.

This means that worlds where Fred is only a little taller don't count as A-worlds. So does "Fred is taller" expresses a proposition here which is not true at worlds where Fred is only slightly taller? That sounds odd.

> **48)** If there had been people in the mall when the fire broke out [on Christmas day], hundreds would have died.

Presumably worlds where only one or two people are in the mall don't count as A-worlds here. So does "there are people in the mall" now express a proposition which is not true at worlds where there are one or two people in the mall? Again, odd.

Yesterday I repaired my bike so that I could ride to the university with it today.

> **49)** If I hadn't repaired my bike yesterday, I would have walked to the university today.

But suppose unbeknownst to me, I very narrowly escaped a fatal heart attack yesterday shortly before repairing the bike. Then arguably the closest worlds at which I didn't repair the bike are worlds where I am dead. But they don't count. The antecedent picks out worlds where I do something else, not worlds where I'm dead. But isn't "I don't repair my bike" simply the negation of "I repair my bike" and thus true at worlds where I'm dead?

> **50)** If Hurricane Katrina hadn't hit the town with 200 km/h, completely destroying our house, we would be at home now, watching TV.

Here the worlds we consider are not worlds where the antecedent is only slightly false – where, say, Katrina hits the town with 201 km/h, or where it only partly destroys the house. Rather, we consider worlds where it didn't hit the town at all, even though these worlds are probably not closer than the others. So does "Hurricane Katrina didn't hit the town with 200 km/h..." here express something that entails that it didn't hit the town at all?

Similarly:

> **51)** if Hitler hadn't come to power in 1933, Hempel wouldn't have emigrated to the US.

Again, this is probably false if worlds where Hitler came to power in 1934 qualify as A-worlds. So does "Hitler didn't come to power in 1933" here express something that entails he didn't come to power at all?

This kind of case is important for Lewis's counterfactual analysis of causation, where "C causes E" is analysed roughly as "$\neg C \boxdot\!\!\rightarrow \neg E$": "When asked to suppose counterfactually that C does not occur, we do not really look for the very closest possible world where C's conditions of occurrence are not quite satisfied. Rather, we imagine that C is completely and cleanly excised from history, leaving behind no fragment or approximation of itself". ("Causation as influence", p.190[4])

4 compare Lewis, *Philosophical Papers 2*, p.211, and Bennett, *Events and their names*, p.55f.

Where does this rule come from? Doesn't Lewis's theory of counterfactuals tell us, on the contrary, that we *have* to look at the very closest possible world where C's conditions of occurrence are not quite satisfied? Why ignore these worlds and look at others that are more remote?

It would be nice if worlds where C isn't "completely and cleanly excised" were simply not worlds where the antecedent $\neg C$ is true in the first place. But it is hard to believe that the antecedent sentences really does express the strengthened proposition in all the cases we've considered.

It would be better to have a pragmatic explanation of what's going on. Compare implicature:

52) Fred had beer or wine.

53) Fred ate some of the cookies.

54) We went to the train station yesterday.

All these would usually be taken to say something stronger than they actually do: that Fred did not have beer *and* wine, that he did not eat *all* the cookies, that we didn't go a lot futher than the train station. The standard explanation is that uttering the sentences would flout conversational rules if one didn't believe those stronger propositions to obtain.

Sadly, I don't know how to explain the unexpected antecedent strength witnessed above in a parallel fashion.

To see the problems, note first that strengthening the antecedent of a conditional does not strengthen the conditional. If the rule of strengthening the antecedent ($A \boxrightarrow C \models A \wedge B \boxrightarrow C$) is valid, the opposite is true. Most likely, there is no general logical relation between a conditional with strengthened antecedent and the original conditional. This makes the present case quite unlike all paradigmatic cases of conversational implicature.

Another (related) disanalogy is that in all the examples considered above, something true is conveyed by saying something that appears to be false on a strict interpretation. Pragmatic phenomena of this kind exist – irony, metaphor, meta-linguistic negation –, but counterfactuals do not fit those templates.

To get around the problems, one might suggest that the relevant antecedent proposition for evaluating counterfactuals is not the proposition expressed by the antecedent but the proposition *implicated* by the antecedent in the utterance context.

This seems initially attractive because in many of the examples, uttering just the antecedent also has the relevant implicature. "Hitler didn't come to power in 1933" can implicate that Hitler didn't come to power at all; "the coin didn't always land heads" can implicate that it didn't always land tails either.

However, these implicatures seem weaker. Also, rules of conversation don't apply to sentences which, like the antecedent of a counterfactual conditional, aren't asserted.

Nevertheless, it seems clear that antecedent strengthening as in (47) – (51) is to be explained pragmatically: we strengthen the antecedent because otherwise what is said would be obviously false.

5 Ignored A-worlds

Some more cases of antecedent strengthening that might require a special treatment.

Suppose I tossed a fair coin each morning for the last 10 days, and it always landed heads.

> **55)** If the coin hadn't always landed heads, people would have found it easier to believe that it is fair.

For whatever reason, we here seem to consider various ways Set aside the fact that on Lewis's standards, the coin would have landed tails once, namely yesterday. This seems clearly false. We spread the antecedent across all other ways the coin could have landed.

So far, the case fits the spreading pattern we've observed before. But notice that we don't really consider *all* other ways the coin could have landed: we don't consider worlds where the coin lands all tails. But arguably, this also isn't like the other strengthening cases. We wouldn't say here, as we would there, "what I strictly meant was that if Φ ...", where Φ is stronger than the actual antecedent. Rather, we'd weaken the consequent: "what I strictly meant was that it would *very likely* have happened that ...". But the difference is subtle.

More examples. It is acceptable to drop the "probably" in (46):

> **56)** If lots of big miracles had occurred yesterday, things would be different now.

Here we ignore the odd possibility that the later miracles undo all effects of the earlier miracles.

Again,

> **57)** if I dropped the plate, it would fall to the floor;

even though it is compatible with the laws of quantum mechanics that the plate would fall sideways. We ignore this odd and unlikely possibility.

What could explain these facts?

One might suspect that the occurrence of events with low objective probability detracts from similarity: extremely improbable events count like miracles; a world where the plate falls sideways isn't strictly a world that violates our laws, but it is so improbable that it counts for almost the same.

But this is wrong. First, the phenomenon also shows up in deterministic worlds where all objective chances are either 1 or 0. Second, the non-ignored worlds are often just as improbable as the ignored worlds: that the plate falls down exactly like *so* is no more probable than that it falls sideways; that the coin lands H-T-H-H-T-T-T-H-T-T is not more probable than that it lands T-T-T-T-T-T-T-T-T-T.

Third, suppose I have a slightly unfair die with an 0.3 chance of landing 6. We don't want to say that

> **58)** # if I had tossed the die (in such-and-such a way), it would have landed 6,

even though this result has a higher chance than all other results. So the mere fact that the 6-world is more probable than the 5-world doesn't make it closer to the actual world.

David Lewis once suggested that what detracts from similarity is not low probability per se, but low probability in combination with being odd, or, as he calls it, remarkable (xxx).

However, it seems that the ignored worlds show up in 'might' (and related) counterfactuals.

59) If the coin hadn't always landed heads, it might have landed all tails.

60) If lots of big miracles had occurred yesterday, things could be just as they actually are.

61) if I dropped the plate, there is a small chance that it would fall sideways.

This should not happen if for whatever reason the worlds count as more distant from actuality.

Compare always – sometimes/once: "Immanuel always got up at 5 in the morning"; that's acceptable, even though "once, when he was sick..." etc. is also true. It's as if universal quantifiers come with a vagueness parameter that can be destroyed by pointing out counterexamples. Also note that here, too, remarkable counterexamples (where Immanuel is sick, very young, dead etc.) are more easily ignored than unremarkable ones (just waking up at 7 twice per week if he was too tired; even if these cases are just as frequent as the remarkable ones).

Indeed, compare "all": "all rooms are occupied"; "everyone is here": we quantify restrictedly, but over what domain? all the occupied rooms/everyone who is here? yes, if context allows and if there is some natural and salient distinction between these items and the others. Interestingly, mentioning somebody else can undermine the "everyone" claim: "but Fred isn't here" – "ok, I meant everyone I expected to come"; and it's odd to say "Fred isn't here, but everyone is here". This is all very similar to would/might.

"I gave presents to all members of my family" – well, except myself.

"I lost my keys; I've already looked everywhere" – well, not on the moon.

"I've told you everything I know about her" – except for all kinds of trivial things, like that she has less than a million hands.

The bible says that every creature not on Noah's ark perished in the flood. – Well, except for the fish, I suppose.

As soon as the counterexamples are mentioned, it gets very hard to go back and assert the universal quantification. Yet the orthodox theory is not that we always quantify unrestrictedly.

So maybe the right thing to say is after all that would goes over every salient world, and that odd ones are just ignored, in the way odd people/rooms/days are often ignored in universal sentences – until somebody explicitly brings them in. We ignore freaks if they make what is asserted clearly false.

One could say that *strictly speaking*, not *everyone* is here unless absolutely everyone is here. But this is a strange notion of strictness that goes beyond ordinary ignorings of pragmatic influences, I think: we really want to say that many (restricted) universal quantifications are, semantically, true, rather than strictly false but pragmatically interpreted to contain some restrictor.

So maybe a better strategy is to treat all these worlds as equidistant, and weaken the semantics of 'would' counterfactuals: $A \mathrel{\Box\!\!\rightarrow} C$ is true iff among the closest A-worlds, *sufficiently many normal ones* are C-worlds. This isn't entirely artificial, as arguably the same applies to other modal statements, plural descriptions and ordinary universal quantifiers:

62) Killing innocent people is not acceptable [except if you know that unless you kill innocent people today, millions will be tortured to death tomorrow].

63) The kids that terrorized our neighbourhood last year are now terrorizing O'Connor [except for one of the kids who has moved to Sydney].

64) All ravens are black [except for albino ravens, and painted ravens].

Note that pointing out the counterexamples in advance makes these sentences unassertable, just like relevant might-counterfactuals make the corresponding would-counterfactuals unassertable ("might fall sideways, but would not"). A good question is why the standards for "might"/"some" are stricter than for "would"/"all"

John Hawthorne (2005) complains that this account makes the logic of counterfactuals *irregular*, that is, that it breaks the following inference rule:

$\Box(B_1 \wedge B_2 \rightarrow C)$
$A \mathrel{\Box\!\!\rightarrow} B_1$
$A \mathrel{\Box\!\!\rightarrow} B_2$

$A \mathrel{\Box\!\!\rightarrow} C$.

I'm not too worried about that. Also, one can keep the logic simple by assuming regularity and saying that *strictly speaking*, (55), (56) and (57) are false, but *almost true*, just as we keep quantificational logic simple by saying that strictly speaking, (64) is false, but almost true, if there are a few non-black freak ravens.

At any rate, the current proposal still presupposes a distinction between *normal* and *non-normal* A-worlds; dropping this clause, we'd only require that *sufficiently many* of the closest A-worlds are C-worlds, and this will presumably not do. For one, it is hard to count the relevant A-worlds; there are infinitely many nomologically possible ways the plate could fall downwards, and infinitely many ways it could go sideways. Moreover, it really seems that remarkable cases are much more easily ignored than unremarkable ones:

65) if I tossed this coin 10 times, it would not land all heads

sounds more acceptable than

66) if I tossed this coin 10 times, it would not land H-T-H-H-T-T-T-H-T-T.

(Compare (62): it's the bizarreness of the counter-example, not so much its rareness that makes (62) acceptable.)

So whether we count remarkable worlds as less similar to actuality or whether we count them as more easily ignored, we do need the distinction.

Lewis introduced remarkable events as detracting from closeness in a slightly different context. He noticed that his similarity standards quoted above do not work for indeterministic worlds. The problem is that in an indeterministic world, it doesn't take miracles (violations of laws) to alter the course of history. In particular, it doesn't take a big miracle to hide all traces of a counterfactual event – say, a stone throwing or a button pressing – shortly after its occurrence and thereby to secure perfect match between the counterfactual future and the actual

future. Hence Lewis says that such worlds where there's a "remarkable way in which the chance outcomes seem to conspire to produce a pattern" (Lewis 1979: p.60) should count like worlds with real miracles.

Our observations in the preceding sections have led to another reason for disqualifying worlds with remarkable patterns, a reason that has nothing to do with chance and is largely independent of Lewis's similarity standards. Even in a completely deterministic world, it is possible that a coin lands heads a million times in a row, or that all the gas molecules in a room suddenly move into one corner, or that a brick thrown at a window passes smoothly through the glass (xxx is that last example from Bennett?). Yet normally we don't consider such worlds when we evaluate counterfactuals, even if they are worlds where the antecedent is true and nothing else we've said rules them out.

Now we should ask a few questions: 1. what should we understand by "remarkable" for it to do the job? 2. Isn't it circular to use a response-dependent concept in an analysis of counterfactuals? 3. Doesn't this introduce a strong relativism into the analysis, as what is remarkable for me in context x may well be unremarkable for you in context y?

Right now, I don't have much to say on the first two questions (except noticing that "unremarkable" seems to be closely related to "random", and there are promising attempts to define the latter). As to the third question, the relativism could be avoided by rigidification, but maybe we should rather accept it. Suppose we just witnessed an unremarkable series of a million coin tosses and wrote down the pattern. Now we toss the coin again a million times. In this context, it would be quite remarkable if the same (previously unremarkable) pattern shows up again, and it is OK to say, beforehand, "if we'd throw it again a million times, we would get a different pattern". The same holds if we've only imagined the previous series. It is undeniable that counterfactuals are highly context-dependent, and a dependence on what is regarded as remarkable by the relevant subjects in the context could go a small step towards an explanation.

John Hawthorne (2005) raises four objections to the proposal that remarkable patterns detract from closeness. I will briefly comment on three of them (the fourth one is closely related to the third, and answered in a similar way).

Hawthorne's first objection is this: suppose in some counterfactual situation A, a certain remarkable outcome C – a monkey writing a novel – is quite probable. In this case, $A \mathrel{\Box\!\!\rightarrow} \neg C$ is false. But in an indeterministic world, every probable event is part of a world that, considered as a whole, is extremely improbable. So the C-worlds are all remarkable and improbable, and if that makes for remoteness (or for being ignored), $A \mathrel{\Box\!\!\rightarrow} \neg C$ will wrongly come out true.

I don't think this argument works. If something had a high chance of coming about, it isn't really remarkable that it comes about. It's remarkable if a fair coin lands heads a thousand times in a row; but if a biased coin always lands heads, that's not remarkable. Likewise, if an ordinary monkey on a typewriter produces a novel by randomly hitting keys, that is remarkable; it isn't remarkable if that happens with a monkey who is likely to produce a novel.

Hawthorne's second objection is that non-remarkable outcomes can be even less probable than remarkable outcomes (with low probability): that the coin lands H-T-H-H-T-T-T-H-T-T is less probable than that it lands either all heads or all tails. So if the remarkable outcomes are ignored, we have to say that under certain conditions, C_1 would have been more likely than C_2 and yet C_1 would not have occurred while C_2 might have occurred. This Hawthorne finds "very

odd indeed".

However, as Robbie Williams (2007: pp.20–25) shows, there are convincing cases where even $A \boxrightarrow \neg C_1$, $A \boxrightarrow C_2$ and $A \boxrightarrow$ (C_1 is more likely than C_2) are all true. For example (my example), let A be some fair coin landing heads five times in a row, C_1 the coin not landing all heads, and $C_2 = A$, and note that the mere fact that something improbable happens doesn't suddenly make it very probable: even in the relevant A-worlds, C_1 is still more likely than C_2. The constellation Hawthorne finds very odd does look odd, but it is actually quite common.

Hawthorne's third objection is that in a big enough world, remarkable events are bound to happen every now and then. What if there were $2^{10'000'000}$ coin flippers each flipping a coin a million times? Discounting remarkable outcomes, we'd have to say that the first series would not land all heads, the second series would not land all heads, and so on. Hence (by regularity), none of the $2^{10'000'000}$ series would be all heads. But this sounds false.

Clearly, a world with $2^{10'000'000}$ series of coin tosses none of which lands all heads is quite remarkable – like an evening of playing cards during which nobody ever has three aces. By contrast, it isn't as remarkable if some more gerrymandered coin toss pattern doesn't show up, or if nobody ever has an ace together with a 7 and a 4. In general, if C is an interesting pattern then both the low-probability occurrence of C and the low-probability *absence* of C are remarkable.

Still, focussing on any single coin flipper, I'm inclined to agree that if *he* were to toss the coin a million times, it wouldn't land all heads. So regularity fails. As noted above, this is quite acceptable if we can also provide a sense in which "strictly speaking", the coin *might* land all heads, and so in some strict and philosophical sense of $\boxrightarrow$, regularity holds. (This is harder if we follow Lewis and say that remarkable and improbable worlds are more remote than if we say that they aren't more remote, but more easily ignored when quantifying – but then we need another answer to chancy re-convergence.)

Robbie Williams (2007) has proposed another way of dealing with remarkable patterns: he suggests to measure a world's closeness to actuality in part by its degree of *fit* to the actual laws of nature, where, roughly, world w_1 has greater fit than w_2 if what happens at w_1 is more probable according to the actual laws than what happens at w_2.

This is an elegant suggestion, but I think it should be rejected. First, it doesn't help to exclude remarkable worlds in the deterministic cases. More importantly, it appears that the proposal shares some of the problems of simply measuring closeness by probability, like the problem with the unfair dice of which it is still false to say it would land on 6. Moreover, it entails that sometimes, events that have a reasonably high probability under certain conditions would not occur if those conditions obtained, and that sounds false. For example, suppose the actual world fits the laws quite well, but its first 100 years didn't (like the first few tosses of a fair coin sometimes all land heads). What if there had been a doomsday machine attached to a fair coin tossing 100 years after the Big Bang such that on tails, the world would have gotten destroyed? I think it's definitely not true that this coin would have landed heads. It could just as well have landed tails. But the world with the doomsday future fits the laws far less than a world with a future closer to actuality.

6 Similarity versus selection

Another pragmatic factor: the similarity standards, so far held fixed, can shift.

We shouldn't take Lewis's list as a general rule. It is a precisification of counterfactuals for a given philosophical project – the analysis of causation.

So far, I've mostly assumed that even though the 'closeness' or 'similarity' relevant for the evaluation of counterfactuals doesn't perfectly correspond to off-hand judgements about similarity, it is at least somehow related to those judgements, enough to deserve the name 'similarity'. But there are reasons to doubt this.

> **67)** if the USA destroyed its weapons tomorrow, there would be war

is arguably true. So is

> **68)** if all nations destroyed their weapons tomorrow, there would be peace.

But if (67) is uttered after (68), it sounds false:

> **69)** # if all nations destroyed their weapons tomorrow, there would be peace; but if the USA destroyed its weapons tomorrow, there would be war.[5]

The worlds where all nations destroy their weapons are still relevant when we evaluate the second part of (69). In the Lewis-Stalnaker framework, this has to be explained by the context-sensitivity of similarity: in the second part of (69), worlds where all nations dump their weapons count among the closest worlds where the US dump their weapons. In general, previously mentioned A-worlds usually count as among the 'closest' A-worlds.

It is clear that "closeness" and "similarity" are now rather misleading terms: in no ordinary sense does something become more similar to actuality merely in virtue of being mentioned.

Indeed, perhaps there are contexts where the relevant A-worlds fail to contain the actual world even though the actual world is an A-world:

> **70)** Suppose fat people were considered beautiful. If you were as thin as you actually are, you would eat more.[6]

Compare ordinary descriptions: "the Fs are G" is true if all *contextually salient* Fs are G, and contextual salience is partly determined by previous discourse. In a suitable context, "the pigs are sick" can be true even if there are very salient pigs nearby that aren't sick, because the previous discourse makes other, far-away pigs more salient. – Imagine the sentence as an answer to "how are things at home?" If counterfactuals are world descriptors, we should expect similar phenomena here.

I'm not sure there are any clear cases of this kind; (70) could be read as involving a hidden embedding of counterfactuals ("suppose ...; *if this were so, then* ..."). But even if we just want to consider this as a possibility, we would do better to avoid terms like "similarity" and "closeness". So let's rephrase the truth-conditions:

5 The example is due to Heim, it is based on an example from Sobel used by Lewis; see von Fintel, "Counterfactuals in a Dynamic Context".

6 modified from Schlenker, op.cit., p.14

$A \boxrightarrow C$ is true iff C is true in all contextually *selected* A-worlds.

At this stage, one could easily build the truth of A into the selection conditions and treat counterfactuals as strict conditionals after all: $A \boxrightarrow C$ is true iff $A \supset C$ is true in all contextually selected worlds. This is what von Fintel suggests. (But unlike me, he stipulates that the actual world is always selected.) I'm not clear about what difference this makes. (It seems that strengthening the antecedent is valid on von Fintel's account but invalid on mine, so there certainly are differences.) For ordinary descriptors, it would mean that "the pigs are sick" is true iff all *contextually salient objects* are either not pigs or sick. An advantage of this approach is that it makes it easier to state rules for the contextual dynamics of the selected/salient domain. A potential disadvantage is that there might be no single contextually relevant domain at work in all descriptors or counterfactuals: maybe in a series of "the pigs...; the dogs...; the pigs...; the dogs...", there sometimes is no reasonable way of extending the domain to cover the relevant dogs without also covering irrelevant pigs. I don't have a convincing example of this however.

Anyway, we still face the question what, in a given context, counts as a 'selected' (A-)world. We know that normally, whatever was previously selected in a discourse remains selected. We also know that normally, worlds where the past is very different from how it actually is are not selected. Many of the arguments in the previous sections can be used to defend more specific proposals, like Lewis's, from apparent counter-examples. But Lewis's proposal is to static: we know that the standards greatly vary with context and are largely determined by pragmatic considerations about what would be reasonable to say.

7 Context-sensitivity

I've already mentioned, in passing, the strong context-dependence of counterfactuals. It is worth having a closer look at this phenomenon.

I arrived in Australia in late October, when it was already quite hot.

71) If I had arrived earlier, I would have noticed that it's not always hot here.

I went by plane from London to Sydney, from where I took the train to Canberra. In between, I spent a night at a youth hostel in Sydney. I didn't get to see much of Sydney because my flight arrived in the late afternoon and my train left very early in the morning.

72) If I had arrived earlier, I would have gone to the harbour before going to the youth hostel.

So what would have been the case if I had arrived earlier? Would I have gone to the Sydney harbour and at the same time have noticed that it's not always hot here? No.

Echoing what I said earlier about "early", one might suggest that the meaning of "earlier" varies from context to context: in (71), it means "several months earlier"; in (72), it means "a few hours earlier".

But this is already hard to believe. And the strategy looks rather hopeless if we look at other cases:

73) If Frege were here, he could explain to us what he thought about the relation between concepts and thoughts.

On the other hand,

74) if Frege were here, he would be so old by now that he probably couldn't explain anything at all to us.

There isn't much ambiguity or indeterminacy to locate in the semantics of "Frege is here". Likewise,

75) if the thermometer showed 0 degrees Celsius, we'd have to heat;

on the other hand,

76) if the thermometer showed 0 degrees Celsius, it would be broken.

Examples abound, and are mostly familiar: what if Verdi and Bizet were compatriots? What if Julius Caesar had been in command in Korea? What if the dinosaurs were still around? What if Clinton was president? And so on.

It is sometimes claimed that there are only a few different kinds of counterfactuals, like forward and backtracking ones, and that one of them is the 'default' one (see e.g. Lewis 1979 34f. and Kment 2006). But that seems unlikely.

Also, we shouldn't focus exclusively on counterfactuals with nomically possible antecedents about local events.

(Intuitively, there are many worlds where I arrived earlier, where Frege is alive today, where Clinton is president, etc. Some of these are irrelevant because too remote: worlds where Australia is uninhabited and ice-covered and I arrived here as part of a scientific expedition, for example. But many others are relevant: worlds where I arrived in much the same way as I did a year earlier or an hour earlier etc.

A counterfactual $A \boxdot\!\!\rightarrow C$ is true iff C holds throughout the relevant A-worlds.

The context of utterance determines which of the pre-selected worlds I'm actually talking about: the ones where I arrived a year earlier, or an hour earlier, etc.)

If there's a threat of misunderstanding, we can, and often do, make the relevant worlds more explicit: "I mean, if I had arrived like several months earlier"; "I mean, if the dinosaurs were alive but we would also be around with our cities etc.".

At any rate, we don't want a similarity metrics on worlds that tells us for any two worlds which one is closer to actually: whether arriving a minute earlier is closer than arriving a few months earlier. At least, we don't want such a metrics that applies to all counterfactuals in all contexts. And a metrics for a single case in a special contexts isn't very helpful in a general semantics of counterfactuals. (Note that it doesn't look like we can do with a single similarity standard and then claim that counterfactuals always only quantify over a subset of the selected worlds: that gives the wrong results for "might" counterfactuals (and it anyway doesn't sound promising).)

But we can say something about which worlds are acceptable as closest depending on the context and which aren't. For clearly not everything goes. You can't say "If Frege was here, he'd wear silly dresses and give lectures on parapsychology" and, if challanged, explain that you really meant "if Frege was here *and would wear silly dresses and would give lectures on parapsychology, then* ...". You *can't* mean that. Or you can, but that's like meaning "green" by "blue": you can't use the sentence you uttered with this meaning in accordance with our linguistic conventions. (So maybe we should ask: what can you mean and what not?)

8 Selection criteria: why bother?

To my knowledge, nobody cares to figure out what exactly determines salience in the evaluation of definite descriptions – the rule that tells us what, in any given context, is denoted by "the door" or "the pigs". Why should we care about the precise standards for counterfactual selection?

Because there's a threat of primitivism here that doesn't exist for descriptions. Here is John Hawthorne again (op.cit.):

> my own preference is to opt for the most straightforward versioin of Robert Stalnaker's semantics for counterfactuals in which, for any possibility that P, and any world w, there is a unique closest world w where P. I realize, of course, that this is to give up altogether on the Lewisian idea of analyzing counterfactual closeness in terms of similarity, and to give up on a thesis of Humean Supervenience [...]. It is also to give up on all neo-verificationist analyses of counterfactual discourse, since the closeness relation between worlds and the counterfactual operator on propositions form a family into which there is no entering reductive wedge.

Note how silly this would be for descriptions: it would amount to saying that salience is a metaphysically and conceptually primitive relation between utterance contexts and objects, not determined by what other properties the contexts and objects have. *Of course* it is ordinary facts about the door and the context that makes my current utterance of "the door" denote this door. If two doors are exactly alike in their intrinsic properties and ordinary relations to us, it would be crazy to assume that nevertheless, "the door" determinately denotes exactly one of them, and that it is an inscrutable metaphysical secret which one it is.

I'm inclined to say the same about counterfactuals: Of course it's the properties of possible worlds and utterance contexts that determine which worlds qualify for evaluating a given counterfactual. If two worlds are intrinsically exactly alike, it is crazy to assume (like Hawthorne) that only one of them gets selected, and that it's a primitive secret which one it is.