Logique & Analyse **239** (2017), 317-335

HOW TO REJECT A COUNTERFACTUAL

Vittorio Morato

Abstract

According to D. K. Lewis (1973), would-couterfactuals and might-counterfactuals are duals. From this, it follows that the negation of a would-counterfactual is equivalent to the corresponding "might-not"-counterfactual and that the negation of a might-counterfactual is equivalent to the corresponding "would-not"- counterfactual. There are cases, however, where we seem to be entitled to accept the would-counterfactual *and* we are also equally entitled to accept the corresponding might-not-counterfactual and cases where we seem to be entitled to accept the might-counterfactual *without* being equally entitled to reject the corresponding would-not-counterfactual. In this paper, I will show that a distinction between two types of rejections for counterfactuals (p-rejection and s-rejection) and the recognition that might-not-counterfactuals may play the role of p-rejections (by an application to counterfactuals of the Lewisian approach to conversational scores) could explain why the problematic cases should not be seen as cases where the duality of would- and might-counterfactuals fails.

1. Introduction

It is usual to distinguish two kinds of counterfactuals. On the one hand, we have the "would-counterfactuals" ("would-cfs", from now on) like the very famous:

(1) If kangaroos had no tail, they *would* topple over;[1]

on the other, we have the "might-counterfactuals" ("might-cfs", from now on) like:

(2) If kangaroos had no tail, they *might* topple over.

The distinction between would-cfs and might-cfs may be and it actually has been contested; in particular, it may be and it actually has been contested that might-cfs form a semantically primitive type of conditional.[2]

[1] See D. K. Lewis 1973, p. 1.

[2] Robert Stalnaker, for example, has defended the view that might-cfs are really would-cfs in the scope of a possibility operator whose interpretation might be epistemic, but also

doi: 10.2143/LEA.239.0.3237157.

The distinction, however, is part of the standard, *Lewisian* picture of counterfactual semantics.

In D. K. Lewis 1973 the relations between would-cfs and might-cfs are regulated by the *Duality Thesis*, according to which might-cfs and would-cfs behave as *duals*:

(DT) $(\phi \diamond\!\!\rightarrow \psi) =_{df.} \neg(\phi \mathrel{\Box\!\!\rightarrow} \neg\psi)$

(DT) is a quite natural thesis to have within a Lewisian framework. For Lewis, would-cfs are expressed by $\Box\!\!\rightarrow$, and $\Box\!\!\rightarrow$ is basically treated as an operator where the antecedent acts as a (variably strict) *necessity operator* on its consequent.[3] The truth conditions of a formula like $\phi \mathrel{\Box\!\!\rightarrow} \psi$ are, in fact, the following:

> (L-$\Box\!\!\rightarrow$) $\phi \mathrel{\Box\!\!\rightarrow} \psi$ is (non-vacously) true at i iff there is a ϕ-world k in S_i such that, *for any world* j, if $j \leq_i k$, then $\phi \rightarrow \psi$ holds at j,

where S_i is the set of worlds accessible from i, $\leq_i$ expresses the relation of comparative similarity (in this case, with respect to i) and $j \leq_i k$ means that world j is at least as similar to i (the world of evaluation) as it is k. Roughly, the idea behind (L-$\Box\!\!\rightarrow$) is that a would-cf is (non-vacously) true in case the antecedent is true in at least a world k and the consequent is true in *every antecedent world* (a world where the antecedent is true) as similar to the the world of evaluation (i.e., i) as it is k.[4]

It is natural to expect that, at least in principle, to this peculiar necessity operator ($\Box\!\!\rightarrow$) there corresponds an equally peculiar possibility operator. This is indeed the case and Lewis chooses to formalize this operator by means of $\diamond\!\!\rightarrow$. $\diamond\!\!\rightarrow$, in analogy with $\Box\!\!\rightarrow$, may be seen as an operator where the antecedent acts as a possibility operator on its consequent and whose truth conditions are thus the following:

> (L-$\diamond\!\!\rightarrow$) $\phi \diamond\!\!\rightarrow \psi$ is (non-vacuously) true at i iff there is ϕ-world k in S_i such that *there is at least one world* j such that $j \leq_i k$ and ψ and ϕ are true in j.[5]

Finally, there is a (quite controversial) linguistic hypothesis: for Lewis, as $\Box\!\!\rightarrow$ expresses natural language would-cfs, $\diamond\!\!\rightarrow$ expresses natural language might-cfs.

non-epistemic; cf. Stalnaker 1984, 145. Keith De Rose has defended a view – actually a variant of Stalnaker's – according to which might-cfs are really would-cfs in the scope of an epistemic operator; cf. De Rose 1999, 389.

[3] For this way of presenting the Lewisian semantics of $\Box\!\!\rightarrow$, cf. Stalnaker 1978, 93.

[4] The truth conditions in terms of comparative similarity are given at page 48 of D. K. Lewis 1973; they are equivalent to the truth conditions given in terms of spheres given at page 16. If we have the limit assumption, but we drop the uniqueness assumption, a simplifified formulation of (L-$\Box\!\!\rightarrow$) would be the following: $\phi \mathrel{\Box\!\!\rightarrow} \psi$ is true at i iff for all closest worlds of i where ϕ is true, ψ is true.

[5] If the antecedent of a might-counterfactual is impossible, then the might-cf is false.

Notice that, according to (DT), $\Box\!\!\rightarrow$ and $\Diamond\!\!\rightarrow$ are duals exactly as the standard modal operators, $\Box$ and $\Diamond$, are duals (at least in their alethic interpretations). Indeed, the analogy is so tight that we can also define modal operators in terms of the counterfactual ones:

- $\Diamond \phi =_{df} \phi \Diamond\!\!\rightarrow \phi$;
- $\Box \phi =_{df} \neg\phi \Box\!\!\rightarrow \phi$.[6]

Like any duality thesis, (DT) gives us some informations about the way in which, given a pair of dual operators, one operator might be negated in terms of the other. From (DT) (and double negation), the following two theses are in fact derivable:

$$\vdash \neg(\phi \Box\!\!\rightarrow \psi) \leftrightarrow (\phi \Diamond\!\!\rightarrow \neg\psi). \quad (3)$$

$$\vdash \neg(\phi \Diamond\!\!\rightarrow \psi) \leftrightarrow (\phi \Box\!\!\rightarrow \neg\psi).[7] \quad (4)$$

According to 3, the negation of a would-cf is equivalent to the corresponding might counterfactual with a negated consequent ("might-not-cf", from now on). According to 4, the negation of a might-cf is equivalent to the corresponding would counterfactual with a negated consequent ("would-not-cf", from now on). Note that, from 3, it also follows that a would-cf is equivalent to the negation of its corresponding might-not-cf and, from 4, it also follows that a might-cf is equivalent to the negation of the corresponding would-not-cf (but this is trivial, because it is, basically, (DT)).

3 and 4 are *logical* equivalences. In order to extend the logical relations among $\Box\!\!\rightarrow$, $\Diamond\!\!\rightarrow$ and $\neg$ to "flesh and bones" states of acceptance or rejection of natural language counterfactuals, we need to assume at least some "bridge principles" connecting logical equivalences with norms of acceptance or rejection.[8] A fairly plausible candidate bridge principle may thus be the following:

[6] The definition of modal operators in terms of counterfactual ones has been recently put at the service of various philosophical projects: see, for example, Williamson 2007, ch. 7 or Kment 2006.

[7] The left-to-right direction of 3 is derived from $\neg(\phi \Box\!\!\rightarrow \psi)$ by an application of DN to obtain $\neg(\phi \Box\!\!\rightarrow \neg\neg\psi)$ and then by applying (DT) to obtain $\phi \Diamond\!\!\rightarrow \psi$. The right-to-left direction is derived from $\phi \Diamond\!\!\rightarrow \psi$ by an application of (DT) to obtain $\neg(\phi \Box\!\!\rightarrow \neg\neg\psi)$ and by DN to obtain $\neg(\phi \Box\!\!\rightarrow \psi)$. The left-to-right direction of 4 is derived from $\neg(\phi \Diamond\!\!\rightarrow \psi)$ by (DT) to obtain $\neg\neg(\phi \Box\!\!\rightarrow \neg\psi)$ and by DN to obtain $(\phi \Box\!\!\rightarrow \neg\psi)$. The right-to-left direction is obtained from $(\phi \Box\!\!\rightarrow \neg\psi)$, by applying DN to obtain $\neg\neg(\phi \Box\!\!\rightarrow \neg\psi$ and by (DT) to obtain $\neg(\phi \Diamond\!\!\rightarrow \psi)$.

[8] As it is well-known, it is not easy to spell out adequately the relations between logic and the normative realm. For the sake of this paper, I will avoid this kind of problems and simply assume that a principle like (L-A) is plausible. To have an idea of the difficulties of spelling out adequate bridge principles between logic and norms of belief or acceptance, see, for example, MacFarlane (2004).

(L-A) If $\vdash \phi \leftrightarrow \psi$, then, if one accepts ϕ, then one is equally entitled to accept ψ and *vice versa*.[9]

Following Stalnaker 1984, 79–81, I take acceptance to be an umbrella term for a family of propositional attitudes (belief, presumption, supposition, etc.) whose members share the following feature: if a subject *s* accepts *p*, then *s* is taking *p* to be true.[10]

Throughout the paper, I will assume a "Fregean" view of the relations between acceptance and rejection according to which to reject a certain statement is to accept the negation of such a statement. Also the relations between states of acceptance and rejection and acts of assertion and denial will be a standard one: I will assume that, if one asserts ϕ, then one is accepting ϕ, if one denies ϕ, then one is rejecting ϕ and thus that, if one asserts $\neg\phi$ one is accepting $\neg\phi$ and thus rejecting ϕ. When I will write that a certain counterfactual is asserted, I will take this to imply that the counterfactual is also accepted (the same goes, *mutatis mutandis*, for denials and rejection).

On the basis of (L-A), 3 and 4 can be projected from the abstract realm of the conterfactual logic into acts of acceptance or rejection of natural language counterfactuals as follows:

(A-Count-1) If one rejects ($\phi \mathrel{\Box\!\!\rightarrow} \psi$), then one is equally entitled to accept ($\phi \mathrel{\Diamond\!\!\rightarrow} \neg\psi$) and *vice versa*.[11]

(A-Count-2) If one rejects ($\phi \mathrel{\Diamond\!\!\rightarrow} \psi$), then one is equally entitled to accept ($\phi \mathrel{\Box\!\!\rightarrow} \neg\psi$) and *vice versa*.[12]

Note, as before, that, from (A-Count-1) and (A-Count-2) it also follows, respectively, that if one accepts a would-cf, then one is equally entitled to reject the corresponding might-not-cf (and *vice versa*), and that, if one accepts a might-cf, then one is equally entitled to reject the corresponding would-not-cf and *vice versa*.[13]

The problem for (DT) and its "pragmatic" counterparts, (A-Count-1)-(A-Count2), is that, for many contingently true counterfactuals, it seems quite easy to construct scenarios that would be appropriately be described as ones where:

[9] Note that the *vice versa* is: if one accept ψ then one is equally entitled to accept ϕ.

[10] Stalnaker 2002, 716 gives the following definition of acceptance: "To accept a proposition is to treat it as true for some reason. One ignores, at least temporarily, and perhaps in a limited context, the possibility that it is false."

[11] Where the *vice versa* is: if one accept ($\phi \mathrel{\Diamond\!\!\rightarrow} \neg\psi$), then one is equally entitled to reject ($\phi \mathrel{\Box\!\!\rightarrow} \psi$).

[12] Where the *vice versa* is: if one accepts ($\phi \mathrel{\Box\!\!\rightarrow} \neg\psi$) then one is equally entitled to reject ($\phi \mathrel{\Diamond\!\!\rightarrow} \psi$).

[13] As before, this is basically a direct consequence of (DT), given that, in general, ($\phi =_{df.} \psi$) entails ($\phi \leftrightarrow \psi$).

- we accept the would-cf *and* we are also equally entitled to accept the corresponding might-not-cf.
- we accept the might-cf *without* being equally entitled to reject the corresponding would-not-cf.

If these scenarios, or their descriptions, are plausible, then there are problems for (DT). On the one hand, if there are cases where I accept a would-cf, but where I would also be entitled to accept the corresponding might-not-cf, then (A-Count-1) is false, because it predicts that the acceptance of a might-not-cf entitles one to reject the corresponding would-cf. But given (L-A), if (A-Count-1) is false, also 3 is false and thus (DT) is false. On the other, if there are cases where I am entitled to accept a might-cf, but not in the position to reject its corresponding would-not-cf, then (A-Count-2) is false, because it predicts that the acceptance of a migh-cf entitles one to reject or deny the corresponding would-not-cf. But, again, given (L-A), if (A-Count-1) is false, also 4 would be false, and thus (DT) too.

The falsity of (DT) would have quite important, if not devastating, consequences for a Lewisian treatment of counterfactuals. We should be clear, however, in what sense the cases that I am going to present would constitute a "failure" of (DT).

Surely, (DT) would not fail as a principle regulating the relations between $\Box\!\!\rightarrow$ and $\Diamond\!\!\rightarrow$, the formal counterparts of would-cfs and might-cfs. As a principle about $\Box\!\!\rightarrow$ and $\Diamond\!\!\rightarrow$, (DT) is unassailable, because it follows from the truth conditions assigned to these operators. The relevant sense in which (DT) fails would be in its capacity, via (L-A), to be a principle capable of describing and predicting our patterns of acceptance and rejection of natural language might-cfs and would-cfs. In this sense, a failure of (DT) would mean that might-cfs and would-cfs, as they are expressed in natural language, are not really duals. But if natural language counterfactuals are not duals, then their truth conditions are not those of $\Box\!\!\rightarrow$ and $\Diamond\!\!\rightarrow$. This could mean that $\Diamond\!\!\rightarrow$ is not the formal counterpart of might-cfs *or* that $\Box\!\!\rightarrow$ is not the formal counterpart of would-cfs. The first option may correspond to the view that $\Diamond\!\!\rightarrow$ does not capture the truth conditions of might-cfs while $\Box\!\!\rightarrow$ correctly captures the truth conditions of would-cfs. One way in which this view could be specified is by saying that natural language might-cfs should be treated not as primitive kinds of counterfactual conditionals of the "might"-kind, but as more complex expressions.[14] The second option is more radical, at least from the standpoint of a Lewisian approach, in that it amounts to the view that $\Box\!\!\rightarrow$ does not capture the truth conditions of

[14] As said above, this view has been variously defended by Stalnaker (1984) and De Rose (1999).

would-cfs. One way in which this view could be specified is by saying that would-cfs should not be treated as (variably strict) necessity operators, namely as universal quantifiers on a set of "closest" worlds.[15] My aim in this paper is to defend (DT), by claiming that the incriminated scenarios are not necessarily to be described as ones where 3 or 4, and thus (DT), fail. The tendency of a speaker *s* to accept a might-not counterfactual in the face of a previous acceptance of a corresponding would-cf or the tendency of a speaker *s* to accept a might-cf without an equal tendency to reject the corresponding would-not-cf do not necessarily signal that *s* is contradicting herself or that *s* now consider false what she has previously accepted. The situation might also be described as one where a certain *conversational negotiation* is taking place. Someone accepting a might-not-cf as a reaction to a previously accepted would-cf may simply be seen as someone who has decided, in the course of a conversation, to be more *permissive* in the evaluation of the counterfactual. Being more permissive in the evaluation of a counterfactual is to evaluate it on the basis of a more "liberal" similarity function. Someone recognising that the original counterfactual has been evaluated with respect to a wrong similarity measure should not be seen as someone recognising that, with respect to the original similarity measure, the counterfactual was false.

In this paper, I will show that, quite ironically, all these phenomena can be easily dealt with using Lewis's theory of conversational scores presented in D. K. Lewis 1979b. Even though Lewis did not mention explicitly counterfactuals in this work, the approach to conversational dynamics developed in that article fits perfectly with the cases that we are going to discuss. My hypothesis will be that, in the course of a conversation, might-not-cfs (asserted as a reaction of a previously accepted or asserted would-cf) or might-cfs (asserted as a reaction to a previously accepted or asserted would-not-cf) can have the role of *conversational shifters* with the effect of raising the conversational score of a given would-cf to a more liberal or permissive level. A might-not-cf, or a might-cf, could then be used also to "pragmatically reject" a would-cf without this implying that the would-cf has been, for this reason, "semantically rejected". (DT), through (A-Count-1) and (A-Count-2), only regulates cases of semantic rejection, namely those cases where a counterfactual is taken to be plainly false. A might-not-cf s-rejects a would-cf in case the might-not-cf is true if and only if the would-cf is false, *with respect to the same similarity measure*.

[15] The view that would-cfs should not be treated as peculiar necessity operators is shared both by those believing that would-cf are truth conditional (like Stalnaker (1968), according to which would-cf are statements about a single possible world) and by those believing that would-cfs are not truth conditional statements (like Edgington (1994) or Barnett (2009), according to which would-cfs are acts of stating within a supposition).

A case where a would-cf and the corresponding might-not-cf are accepted is not necessarily to be seen as a case where the would-cf is semantically rejected by the might-not-cf. The two novel notions of *pragmatic rejection* and *semantic rejection* of a counterfactual, and their relations, will be discussed below.

2. The scenarios

Consider the following scenario:

> **Airport.** After a desperate run from the other terminal, I arrive too late at the gate and lose my flight.

In such a situation, it seems that I am entitled to assert something like:

(5) Damn! If I had run faster, I would have caught my flight.

Suppose now that a fellow traveller, call her "the sceptic", having heard my utterance of 5, reacts by saying something like the following:

(6) Do not worry too much, if you think it through, [had you run faster] you might have fallen and lost your ticket.

6 seems plausible; after all, it was *not impossible* that I fell, while desperately running to the gate. More importantly, the possibility of me falling is something that I am already recognising, maybe in the back of my mind, while uttering 5. This possibility, however, was not somehow relevant when I have decided to accept 5. This is compatible with the predictions of many cognitive studies on how people think counterfactually: counterfactual thinking is often generated by regrets for actions with a bad outcome. If I have regrets for having lost my flight (definitely, a bad outcome), I will thus tend to imagine a counterfactual situation where I was able to catch my flight.[16]

But, if I accept the sceptic's suggestion, namely, if I accept 6, I should also accept one of its plausible consequences:

(7) If I had run faster, I might not have caught my flight.[17]

But according to 3, 7 is equivalent to the negation of 5. By (L-A) and (A-Count-1), one should thus claim that my entitlement to accept 7 should

[16] See Byrne (2005, p. 45).

[17] I am assuming here that one knows that losing the ticket prevents one, in normal conditions, from taking a flight and some sort of closure principle for $\diamond\!\!\rightarrow$, according to which, given that one knows that γ is entailed by β and accept $\alpha \diamond\!\!\rightarrow \beta$, then one is entitled to accept $\alpha \diamond\!\!\rightarrow \gamma$.

be equal to my entitlement to reject 5 and, on the contrary, that my entitlement to accept 5 should be equal to my entitlement to reject 7.

The problem is that the situation is not easily describable in such a way. On the one hand, my final acceptance of 7 does not seem to be accompanied by an equal tendency to reject what I have previously asserted, and thus accepted, by 5. On the other, my original assertion of 5 does not seem to be accompanied by an equal tendency to reject 6 (or directly 7). If that would have been the case, my reaction to the sceptic would have been similar to the reaction I would have had with someone plainly contradicting what I have just asserted. However, I did not "perceive" the sceptic as someone contradicting me, but rather as someone asking me to be more careful in the evaluation of my original counterfactual statement. So it seems that, in both cases, my patterns of acceptance/rejection of 5, while asserting or accepting 7, and of 7, while asserting or accepting 5, are not correctly predicted by (DT), (L-A), and (A-Count-1).

Airport is not an isolated case. What is really problematic about it is that it seems just an instance of a very wide class of cases. Most (if not all) contingent counterfactuals can be associated to their "sceptical" counterparts, namely to their corresponding might-not-cfs, and for most of them we can imagine scenarios that can be plausibly described as ones where we first seem to accept the former and, later on, we seem to accept the latter. If these cases are described, through a rigid application of (DT), as ones where the final acceptance of the might-not-cf implies the falsity of the original would-cf, then we should conclude that most contingent counterfactuals are false and that we should probably simply retreat from accepting them. If we find the description of **Airport** plausible and apply (DT), we should then endorse a quite radical form of *counterfactual scepticism*.[18]

My view, however, is that **Airport** (and all cases similar to it) should not be described as one where we explicitly contradict ourselves or retreat what we have previously accepted, nor as one where (DT) fails. On the one hand, my original assertion of 5 is not perceived as improper or plainly false, even when I finally decide to accept or assert 7, on the other, my initial assertion of 5 seems to be not incompatible, *mutatis mutandis*, with what, later on, I accept by accepting 7. In section 3, I will present, in a more detailed way, my diagnosis of the situation.

For the moment, I would like to present another structurally similar scenario, where (DT) seems now to predict fairly well our patterns of acceptance or rejection of counterfactuals.

[18] Counterfactual scepticism, the view that most contingent counterfactuals are false is defended in Hajek ms. For Hàjek, at least one of the reasons why most contingent would-cfs are false is because their corresponding migh-not-cfs are true due to indeterminism or indeterminacy.

Money. Yesterday I had no penny left in my pocket and I know it.

In this scenario, it would be perfectly appropriate to assert something like:

(8) If I had looked in my pocket, I would not have found a penny.

In such a case, given our scenario, it is difficult to imagine how I could accept the sceptic's suggestion:

(9) If you think it through, [had you looked in your pocket], you might have found a penny

Given my epistemic status (i.e., my knowledge of not having had a penny yesterday), it is difficult to imagine that I would accept easily (or, at least, so easily as I would accept 7) something like 9. If I really know that yesterday I had not a penny, then it would be improper (to say the least) for me to accept that I might have found a penny in my pocket. Contrary to the airport case, my acceptance of 8 seems already to exclude my acceptance of 9 and my eventual acceptance of 9 probably would imply a retreat from my previous assertion of 8.[19] In such a case thus, (DT) seems to predict the right reactions of speakers.

Money is the scenario that Lewis himself used to motivate (DT) and to defend his conception of might-cfs.[20] However, as Keith De Rose (1994) as shown, a little variation in the scenario is enough to restore the plausibility of the sceptical reaction. Consider this modified scenario:

Money*. Yesterday I had no penny in my pocket *and I do not know it*

In such a situation, I would *not* be in the position to accept or assert neither of the following two would-cfs:

(12) If I had looked in my pocket, I would not have found a penny;

(13) If I had looked in my pocket, I would have found a penny.

If I do not know whether yesterday I had any penny in my pocket, it would be improper for me to accept that, had I looked, I would have found one, but it would also be improper to accept that, had I looked, I would *not* have found one. Given my epistemic status, however, it was open for me the

[19] As an evidence for this, notice that in the **Airport** scenario, I may say something like:

(10) If I had run faster, I would have caught my flight. Of course, I might have felt, but let us not consider this eventuality;

without this sounding improper or contradictory. On the contrary, in the **Money** scenario, the following would sound improper and almost contradictory or, at least, more improper or contradictory than 10:

(11) If I had looked in my pocket, I would not have found a penny. Of course, I might have found one, but let us not consider this eventuality.

[20] See D. K. Lewis 1973, 81.

possibility of finding some penny yesterday as it was the possibility of not finding some penny yesterday. In such a situation, the sceptic could surely make me accept the following might-cf:

(14) If I had looked in my pocket, I might have found a penny

But (DT) and (A-Count-1) predict that my entitlement to accept 14 should be equal to my entitlement to reject 12. As we have seen, however, we are not in a position to reject 12, even in the case we accept 14, so, again, the predictions of (DT) seems to be wrong in this case.[21]

The difference between **Money** and **Money*** is epistemic in nature, but this does not mean that the possibility of sceptical reactions to contingent would-cfs depends on epistemic features of the scenarios. What the change of epistemic status in **Money*** contributes to is to make some possibilities *salient*. But there are many ways, not only epistemic, to make a possibility salient. For example, some possibilities may become salient in a given scenario because the scenario has to be framed within an indeterministic universe. For example, if you evaluate a would-cf like:

(15) If I had dropped that vase, it would have fallen and broken.

in a scenario where certain interpretations of quantum mechanics hold (for example, one according to which the wave function assigns probability of location), then the possibility for the vase to land safely in the sofa becomes salient and so it seems that, in such a situation, we would be entitled to assert something like:

(16) If I had dropped that vase, it might have landed safely on the sofa.[22]

Surely, even in these cases, we are facing an epistemic phenomenon, because salience is an epistemic notion, but what would not be epistemic are the *reasons* why such possibilities have become salient. A possibility can become salient in a given scenario for non-epistemic reasons.[23]

[21] The case might also be presented as one where I reject both 12 and 13; in such a case, my eventual acceptance of 14 would rightly corresponds, according to (DT), to my original rejection of 14. The problem is that it does not seem plausible to assume that, in a case where I do not know whether yesterday I had some penny in my pocket, I should *reject* 12 and 13. To accept the negation of 12, I should accept that the negation of 12 is false, but this is explicitly excluded by the scenario.

[22] See Hawthorne 2005 for a discussion of this kind of cases. See also K. S. Lewis 2016 where a contextual semantics and pragmatics to would and might-cf is defended which is very similar, at least in spirit, to the proposal of this article.

[23] One might claim that even **Airport** could be seen as an epistemic case: my tendency to accept 7 may be seen as a consequence of the fact that, after having asserted 5, I suddenly realise that *I do not know* that I would surely have reached the gate in time, if only had I run faster. Admittedly, this might be a fair reconstruction of **Airport**, but I would take it with care: on the one hand, we can very often "force" an epistemic reconstruction of a

Note that (DT) seems perfectly restored as soon as we switch from the evaluation of counterfactuals within conversational, contextualised, scenarios to the evaluation of single, sentences. The following sentences, in their natural reading, seem all to be false and rightly perceived almost like contradictions, exactly how (DT) predicts:

(17) If I had run faster, I would have catched my flight but I might not have catched my flight.

(18) If I had looked in my pocket, I would not have found a penny but I might have found a penny.

The suspect is thus that the problems for (DT) coming from **Airport** or **Money*** may depend on the conversational, dynamical elements of these scenarios. The hope is, therefore, to find a way in which these scenarios can be described without necessarily concluding that (DT) fails in them. Let us see how.

3. A diagnosis, and a solution

In this section, I want to propose a diagnosis for the phenomenon under discussion and a related solution, one that is "conservative" with respect to (DT) and, more generally, with respect to a Lewisian treatment of would and might-cfs.

My diagnosis starts from the recognition that *two mechanisms* seems to be at work in the evaluation of a would-cf like $\phi \boxrightarrow \psi$. On the one hand, we have:

(Sel) The **selection** of a relevant class of ϕ-worlds (i.e., the selection of the relevant possibilities);

on the other, we have:

(Det) The **determination** of the truth value of ψ within the class of worlds selected by (Sel).

(Sel) is, basically, a "pragmatic" mechanism. Whether a class of ϕ-worlds is relevant for the evaluation of a counterfactual depends on the notion of

scenario by emphasising its epistemic features (even when such features are simply a by-product of non-epistemic ones), on the other, the evaluation of counterfactuals should not be too strictly related to the epistemic states of the agents. The epistemic status of agents is not relevant in the evaluation of would-counterfactuals, namely in the process of evaluating whether the antecedent variably strictly necessitate the consequent. A would-cf is true not in the case, for the speaker, the antecedent epistemically necessitate the consequent, but in the case the antecedent "metaphysically requires" the consequent. See Barnett 2011 for a defence of this view in the context of a suppositional theory of counterfactuals.

counterfactual closeness and, according to Lewis, counterfactual closeness is to be analysed in terms of similarity among worlds. The determination of similarity among worlds is highly contextual[24], but Lewis believes (at least since D. K. Lewis 1979a) that there is also a *default similarity ordering* that we should follow, unless there are specific reasons to avoid it.

(Det) is instead a mere semantic affair: according to (Det), *once the relevant set of ϕ-worlds has been chosen*, it has to be determined whether ψ is true in all these ϕ-worlds. Note that, in the "official" truth conditions for would-cfs (the ones in terms of spheres at p. 16 or the ones in terms of comparative similarity at p. 49 of D. K. Lewis 1973), there is no explicit mention of something like (Sel). Those truth conditions simply take for granted that the contextual factors determining the similarity measure has been already resolved when the counterfactual is to be semantically evaluated. For a counterfactual to be true, those truth conditions require that, whatever be the choice of the relevant class of ϕ-worlds, within that class, (Det) be satisfied.

While not explicitly part of the official truth conditions for would-cfs, still, the choice of relevant ϕ-worlds through the choice of an appropriate similarity measure seems to be an essential ingredient in the way in which subjects evaluate counterfactuals and manage counterfactual reasonings. If something goes wrong with (Sel), something goes wrong with the overall evaluation of the counterfactual.

In general, we might say that a would-cf is *successfully evaluated* in case both (Sel) and (Det) are successfully "performed". A successful "performance" of (Sel) means that an appropriate set of closest worlds has been selected by the speaker, namely that a relevant set of possibilities has been chosen. A successful "performance" of (Det) means that the speaker is able to correctly determine the truth value of ψ in all chosen ϕ-worlds.[25]

If (Sel) and (Det) determine the conditions for a successful evaluation of counterfactuals, they will also determine the conditions for an *unsuccessful evaluation*. An *unsuccessfully evaluated* counterfactual is one where either (Sel) or (Det), or both, have not successfully been "performed", namely one

[24] Here is a quotation from *Counterfactuals*:

> The truth conditions for counterfactuals are fixed only within rough limits; like the relative importances of respects of comparison that underlie the comparative similarity of worlds, they are a highly volatile matter, *varying with every shift of context and interest*. (D. K. Lewis 1973, 92, my emphasis)

[25] Of course, it is hard to believe that standard speakers of English communicate their opinions about rejections (or about their evaluation) of would-cfs directly in terms of possible worlds (unless they are philosophers) and, at least in this paper, I do not want to take a stand on the issue of the psychological reality of possible worlds semantics. So, take these judgements as idealisations useful only to "dramatise" the point I am making.

where either a non-relevant, contextually inappropriate class of worlds has been chosen or where the truth values of the consequent has not been adequately determined, or both.

An unsuccessful evalutation of a counterfactual is thus one where a speaker makes at least one of these two *errors*:

> (Error-Sel) Errors in the selection of the class of ϕ-worlds.
>
> (Error-Del) Errors in the determination of the truth value of ψ within the class of ϕ-worlds determined by (Sel).

To these two types of errors in the evaluation of would-counterfactuals correspond *two ways of rejecting* would-counterfactuals. If one believes that an error has been done in the evaluation of a proposition, then one typically is entitled to reject the proposition in question. In case a subject *a* believes that a subject *b* (not necessarily distinct from *a*) has made some errors in the evaluation of a certain would-cf, then *a* seems to be in a good (in the best?) position to reject the would-cf. But given (Det) and (Sel), one could then reject a would-cf (i) by pointing out errors in the selection of the relevant class of antecedent-worlds, namely by signalling that an error of type (Error-Sel) has been made, (ii) by pointing out errors in the determination of the truth value of the consequent in the relevant class of antecedent-worlds, namely by signalling that an (Error-Del) has been made, or (iii) by pointing out that both types of error have been made. In the first case, *a* is rejecting *b*'s would-cf by communicating something like "what you have asserted is wrong, because you have been too restrictive (or too liberal) in the selection of the closest worlds", in the second case, *a* is rejecting *b*'s would-cf by communicating something like "what you have asserted is wrong, because, within the class of the worlds you have chosen, not every antecedent-world is also a consequent-world".

On the basis of these considerations, we can now explicitly define two types of rejections for would-cfs:

> **S-rejection:** A counterfactual $\phi \mathrel{\Box\!\!\rightarrow} \psi$ is s-rejected iff it is shown that, while evaluating it, errors of type (Error-Del) have been made, namely iff it is shown that $\phi \mathrel{\Box\!\!\rightarrow} \psi$ is false.
>
> **P-rejection:** A counterfactual $\phi \mathrel{\Box\!\!\rightarrow} \psi$ is p-rejected iff it is shown that, while evaluating it, errors of type (Error-Sel) have been made, namely iff it is shown that an inappropriate comparative similarity system has been chosen.

To s-reject $\phi \mathrel{\Box\!\!\rightarrow} \psi$ one has to show that $\phi \mathrel{\Box\!\!\rightarrow} \psi$ is false, namely that there is a ϕ–world k in S_i and there is world j such that $j \leq_i k$ and $\phi \rightarrow \psi$ is false at j, namely that there is a ϕ-and-$\neg\psi$-world j such that j is at least as similar to i as any ϕ-and-ψ-world. A would-cf is thus s-rejected, if the standard Lewisian truth-conditions are falsified. If one s-rejects a would-cf, then one is accepting the negation of such a would-cf.

P-rejection is a different matter. As I have claimed above, in the "official" truth conditions of a would-cf, the only thing required to S_i is its existence; no reference is made to its appropriateness with respect to the context of utterance. From this, it follows the most relevant feature of p-rejection, at least for my aims: *a p-rejected would-cf is not (necessarily) a false counterfactual*, because the inappropriateness of the similarity system does not bar the would-cf from satisfying the truth conditions stated in (L-□→). One could thus p-rejects a counterfactual without committing herself to the claim that the p-rejected counterfactual is false. One could thus p-rejecting a would-cf without accepting the negation of the would-cf with respect to the original set of relevant possibilities.

We can now use our two novel notions to analyse our problematic scenarios. Both **Airport** and **Money*** are scenarios that can be analysed in terms of p-rejection and thus where the original counterfactuals are p-rejected without being taken to be false. When I finally accept 7, I can be described not as s-rejecting my original utterance of 5, but rather as p-rejecting it. By p-rejecting 5, I am not committed to the claim that 5 is false, but to the claim that the similarity system with respect to which 5 has been evaluated was inappropriate. If this is a fair description of what happens in **Airport**, **Airport** cannot be presented as a scenario where (DT) fails. (DT) is safe, because, in such a scenario, the role of 7 is not that of semantically rejecting 5, but simply that of signalling that a wrong similarity system has been chosen. The final acceptance of 7, the original acceptance of 5 and (DT) are thus compatible. (DT) only predicts that the acceptance of a would-cf is equivalent to the rejection of the corresponding might-not-cf just in case both the would-cf and the might-not-cf are evaluated *along the same similarity measure*. A might-not-cf that p-rejects a previously asserted or accepted would-cf asks instead for a change of similarity measure and should not be seen as a case regulated by (DT) at all. 7 and 5 should not be evaluated along the same similarity measure, so there is no need to apply (DT), but then there is also no violation of (DT).

The same goes for **Money***. When I finally accept 14, I should not be seen as s-rejecting 12 and so not be seen as implying that 12 is false. S-rejection can be done only relatively to the same similarity measure and 12 and 13 are not to be evaluated along the same similarity measure of 14. To p-reject 12 is compatible with a situation where I am not in a position to accept either of 12 or 13. The cause of my incapacity to accept either of 12 or 13 is that I am not in a position to choose an appropriate similarity system to evaluate them. Even in such a case, however, the scenario cannot be presented as one where (DT) fails: (DT) predicts that, if one accepts 14, then one should reject a corresponding would-not-cf that is evaluated along the same similarity measure, but this is not the case with 12. In **Money***, when I do not know whether to accept 12 or 13, I am not able to choose

any similarity measure. But when I accept 14, I have chosen a certain similarity measure. The non-acceptance of 12 and 13 and the acceptance of 14 are thus done with respect to different sets of relevant possibilities. As it was the case with **Airport**, (DT) is safe, because we have a scenario where (DT) is not to be used.

My claim is thus that the recognition of p-rejection is a way to save a Lewisian treatment of counterfactuals from "sceptical" scenarios such as those represented in **Airport** and **Money***. The recognition of p-rejection is not particularly philosophically committing: as we have seen, its postulation is quite a natural consequence of the recognition of two fundamental mechanisms at work in our evaluation of a would-cf, namely (Sel) and (Det).

This picture presupposes that the role of might-not-cfs is not always that of playing the role of s-rejections of their corresponding would-cf, namely that a might-cf should not be seen as always implying the falsity of the corresponding would-cf.

My hypothesis is that might-not-cfs may be asserted (as a reaction to a previous acceptance or assertion of a would-cf) with the intention of signalling the inappropriateness of the would-cf. In typical cases, such as **Airport**, the speaker asserting a might-not-cf is simply asking her interlocutor to be more permissive, namely to evaluate the counterfactual according to a more "liberal" similarity measure.

4. Counterfactual scores

Quite ironically, these conversational processes could be modelled by adapting some tools developed by Lewis himself in a work not explicitly related with the semantics of counterfactuals. In "Scorekeeping in a language game" (1979b), Lewis develops a general theory of how "conversational scores" evolve in the course of a conversation.[26] A conversational score (with respect to a conversation) is, basically, a set consisting of various (abstract) materials such as sets of presupposed propositions, rankings of comparative salience, boundaries between permissible and impermissible courses of action, etc. According to Lewis, at the beginning of a conversation, the various components of a conversational score are set to certain values. It may happen, however, that, as time goes by and the conversation proceeds, the value of some components change. These changes in the conversational score are governed by a "rule of accomodation". The general

[26] Lewis's approach to conversational dynamics has already been used to analyse counterfactuals in Gillies 2007. However, Gillies defends a non-Lewisian theory of counterfactuals according to which would-cfs are strict conditionals.

scheme of the rule of accommodation for a conversational score is the following:

> **Rule of accomodation for scores:** If at time *t* something is said that requires component s_n of conversational score to have a value in the range *r* if what is said is to be true, or otherwise acceptable; and if s_n does not have a value in the range *r* just before *t*; and if such-and-such further conditions hold; then at *t* the score-component s_n takes some value in the range *r*.[27]

Such a schematic rule accounts for the fact that, at least in typical cases, a conversation is a situation where all the participants have the tendency to make everything that is being said as much "acceptable" as possible, provided that none of the participants has something to object.

A typical case in which this happens is the interpretation of definite descriptions.

Consider the following scenario:

> **Cat:** In the room there is a cat and, at a certain point, this cat jumped inside the carton.

In this scenario, imagine that someone asserts something like the following:

> (19) The cat is in the carton. The cat will never meet our other cat, because our other cat lives in New Zealand. Our New Zealand cat lives with the Cresswells. And there he'll stay, because Miriam would be sad if the cat went away.[28]

19 seems perfectly plausible, but to properly interpret it, the denotation of the last occurrence of "the cat" has not to be the one of the first occurrence. The plausibility of 19 may be seen both as a counterexample to the view that the denotation of a definite description "the *F*" is the one and only *F* existing, but also to the view that the denotation of "the *F*" is the one and only *F* existing relative to a contextually determined domain of discourse (because, in 19, the context of utterance is the same). To properly interpret 19 one has to assume that the denotation of "the cat" is the most (contextually determined) salient cat and that what counts as the most salient cat is changed in the course of the conversation. This process of salience shifting is a case of score accommodation perfectly predicted by our general schematic rule. To make 19 acceptable, one has to imagine a change in the rankings of salience that make the cat inside the carton less salient than the cat living in New Zealand.

The hypothesis I am defending here is that the role of might-not-cfs in a conversation (unless this is explicitly contested by one of the participants)

[27] D. K. Lewis 1979b, 347.

[28] This example is at p. 348 of D. K. Lewis 1979b.

is, *at least in some cases*, just that of changing (one of the values of) the conversational score. In the case of the evaluation of counterfactuals, we have to assume that one of the elements of the score is the ranking of possibilities; such a ranking will give us some information about the salience of a possibility with respect to what is said. In some cases, the role of a might-not-cf is that of changing such ranking, by changing the salience of a given possibility.

Take the case of **Airport**. At the time of my uttering 5, the ranking of the possibility of me falling while running to the gate is low. This is somewhat reasonable: 5 is asserted after I have lost my flight in a situation where I am considering all alternative courses of actions that would have made possible for me to catch my flight. The sceptic's reaction, counting as a p-rejection of my original would-cf, has the effect of making the possibility mentioned in 7 salient to the conversation. My acceptance of 7 is thus an effect of the rule of accommodation for scores applied to the rankings of possibilities. Given that the observation of the sceptic is not taken to be unacceptable (in such a case I would probably have the tendency to react immediately to restore the original ranking), the ranking is changed in order to reflect, to accommodate, this change of salience. The effect of 7 is thus that of shifting the conversational score through a reorganisation in the ranking of salient possibilities used to evaluate the would-cf.

Take the case of **Money***. In such a case, the scenario could be described as one where the values in the conversational score have not been explicitly set, given that nothing has been asserted and thus accepted. We could imagine, however, that, given the epistemic situation of the speaker, the ranking of possibilities is such that neither the assertion of 12 nor 13 is taken to be acceptable. The first move in the conversation, namely the assertion of 14, sets an explicit value on which the speakers accomodate, given that what is said seems to be acceptable, after all.

What is important to note is that, through this analysis, we are now able to see why, both in the case of **Airport** and **Money***, the final acceptance of 7 and 14 was not accompanied by an equal tendency to, respectively, reject 5 or 12. *With respect to the original conversational scores*, the assertion of 5 and the non-rejection of 12 can now be explained as perfectly sensible linguistic or cognitive moves. Our tendencies to accept or reject should then be seen as relativized to a certain state of the conversational score and a principle like (DT), or (A-Count-1) or (A-Count-2), are violated only in the case where, with respect to a certain state of the conversational score, both a would-cf or its corresponding might-not-cf are accepted or asserted. Both **Airport** and **Money*** are thus not cases where (DT), or (A-Count-1) or (ACount-2) are violated. P-rejection of a would-cf, done by means of the corresponding might-not-cf could be seen, in effect, as nothing else than a request to change the conversational score. Therefore, a case of

a p-rejected would-cf by means of the corresponding might-not-cf cannot count as a case of violation of (DT) and its pragmatic counterparts.

5. Conclusion

In this paper, I have defended the view that scenarios where we do not have an equal tendency to assert a would-cf and reject the corresponding might-not-cf or where we have the tendency to assert a might-cf without an equal tendency to reject the corresponding would-not-cf should not be seen as cases where a fundamental principle of the Lewisian approach to counterfactuals, namely (DT), fails.

The distinction between p-rejection and s-rejection, the recognition that might-not-cfs may play the role of p-rejections of previously accepted or asserted corresponding would-cfs and, finally, the view that this role can be adequately modelled by an application to counterfactuals of the Lewisian approach to conversational dynamics could effectively explain why cases like **Airport** or **Money*** should not be seen as cases where (DT) or its pragmatic counterparts (A-Count-1) or (A-Count-2) fail.

References

[1] Barnett, D. (2009). The myth of the categorical counterfactual. *Philosophical Studies*, *144*, 281–296.
[2] Barnett, D. (2011). Counterfactual entailment. *Proceedings of the Aristotelian society*, *CXII*, 73–97.
[3] Byrne, R. M. J. (2005). *The rational imagination*. The MIT Press.
[4] De Rose, K. (1994). Lewis on 'might' and 'would' conditionals. *Canadian Journal of Philosophy*, *24*, 413–418.
[5] De Rose, K. (1999). Can it be that it would have been even though it might not have been? *Philosophical Perspectives*, *13*, 385–413.
[6] Edgington, D. (1994). On conditionals. *Mind*, *104*, 235–330.
[7] Gillies, A. S. (2007). Counterfactual scorekeeping. *Linguistics and Philosophy*, *30*, 329–360.
[8] Hajek, A. (ms). *Most counterfactuals are false.* (http://philrsss.anu.edu.au/people-defaults/alanh/papers/MCF.pdf)
[9] Hawthorne, J. (2005). Chance and counterfactuals. *Philosophy and Phenomenological Research*, *LXX*, 396–405.
[10] Kment, B. (2006). Counterfactuals and the analysis of necessity. *Philosophical Perspectives*, *20,* 237–302.
[11] Lewis, D. K. (1973). *Counterfactuals*. Oxford: Blackwell.
[12] Lewis, D. K. (1979a). Counterfactual dependence and time's arrow. *Noûs*, *13*, 455–476.
[13] Lewis, D. K. (1979b). Scorekeeping in a language game. *Journal of Philosophical Logic*, *8,* 339–359.

[14] Lewis, K. S. (2016). Elusive counterfactuals. *Noûs*, *50*, 286–313.
[15] MacFarlane. (2004). *In what sense (if any) is logic normative for thought?* (http://johnmacfarlane.net/normativity_of_logic.pdf)
[16] Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (pp. 98–112). Oxford: Blackwell.
[17] Stalnaker, R. (1978). A defense of conditional excluded middle. In W. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 87–103). Dordrecht: Reidel. Stalnaker, R. (1984). *Inquiry*. The MIT Press.
[18] Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy*, *25*, 701–721.
[19] Williamson, T. (2007). *The philosophy of philosophy*. Oxford: Blackwell.

Vittorio Morato
Department of Philosophy, Sociology,
Education and Applied Psychology University of Padua
vittorio.morato@unipd.it