Forthcoming in *Pacific Philosophical Quarterly*

Lowe's Argument against the Psycho-Neural Token-Identity Thesis

Katarzyna Paprzycka

Abstract. Lowe argues that the mental event token cannot be identical to the complex neural event token for they have different counterfactual properties. If the mental event had not occurred, the behavior would not have ensued, while if the neural event had not occurred, the behavior would have ensued albeit slightly differently. Lowe's argument for the neural counterfactual relies on the possible-world semantics, whose evaluation of such counterfactuals is problematic. His argument for the mental counterfactual relies on a premise that is plausibly false. The arguments presented support other counterfactuals, which are consistent with identity theories.

According to any psycho-neural identity theory, the event token *D* (an agent's decision to raise an arm) is identical with a certain very complex neural event token *N*. Lowe (2006; 2008) argues that if the agent had not decided to raise her arm (where the mental event is an event token), the arm would not have risen. Yet, if a complex neural event token (the potential correlate of the agent's decision) had not occurred, the arm would have risen almost exactly the way it actually did. Since the two event tokens have different counterfactual properties, they cannot be identical.

The argument plays a key role in Lowe's non-Cartesian substance dualism for it demonstrates among others that the self is not a 'mere "shadow" of a substance' (2008, p. 99) but has 'distinctive and independent causal powers' (p. 99). Lowe takes it to underwrite his conception of mental and physical causation. Decisions determine whether and what sort of bodily movements occur (but do not determine their precise physical characteristics) while

neural events determine only the precise parameters of bodily movements (but are not causally responsible for the fact that those movements occur).

The 'neural counterfactual' and the 'mental counterfactual' are established by separate arguments. I consider them in turn and show that Lowe fails to establish the truth of either counterfactual. He uses the standard (Lewisian) possible world semantics to establish the neural counterfactual but fails to recognize that the counterfactual is one of a species of counterfactuals that are problematic for the standard account (§1). His argument for the mental counterfactual, on the other hand, relies on a premise that is quite plausibly false (§2). The minimal result is that Lowe has failed to establish the truth of either counterfactual and so has failed to prove that the token-identity thesis is false. The arguments presented support other counterfactuals, which are consistent with identity theories (§3).

1. What would have happened if not all the neurons had fired the way they did?

Lowe's argument for the neural counterfactual is based solely on the verdict of the standard possible world semantics (§1.1). I show that there are at least two controversies, which are relevant to the evaluation of the counterfactual. In fact, the defender of the identity theory could well appeal, quite independently, to either of these controversies to argue that the verdict is not binding on her (§1.3-§1.4). I begin, however, by sketching some counterexamples (§1.2) – false counterfactuals that share the form of the neural counterfactual, which could be shown to be true on the standard account (as Lowe applies it). I conclude by identifying some contexts in which

we might be willing to accept the neural counterfactual (and similar counterfactuals) as true (§1.5).

1.1. The Argument for the Neural Counterfactual

Consider a complex neural event token *N*, which we can think of as a huge sum of individual neuron firings. *N* causes the agent's arm to rise. What would have happened if *N* had not occurred? Lowe claims:

> If *N* had not occurred, the agent's arm would still have risen in almost exactly the same way as it actually did (2008, p. 105, emphasis removed).

He argues for the truth of the neural counterfactual using Lewisian possible-world semantics (Lewis 1973). In the closest antecedent worlds (where *N* does not occur), there occurs another neural event token *N**, which must be very similar to *N*. This seems to be required by the closeness relation: worlds where *N** occurs are more similar to the actual world than worlds where very many of the relevant neurons fire differently than they actually do. Because of the massive similarity, *N** and *N* will have similar causal powers and so *N** will also result in an arm rising *B**, which will be very similar to the actual arm rising *B*.

Since the token neural event *N* is very complex, it is informative and instructive to cast the neural counterfactual in a way that will reveal its complexity. Let the token neural event *N* be composed of the token events ($N_1, \ldots, N_k$) of neurons $n_1, \ldots, n_k$ firing the way they actually do. Necessarily, *N* occurs if and only if N_1 occurs, . . . , and N_k occurs, i.e. all neurons $n_1, \ldots, n_k$ fire the way they actually do. This means that Lowe's neural counterfactual is equivalent to:

(N#) If not all $N_1, \ldots, N_k$ occurred (if not all neurons had fired the way they actually did), the agent's arm would still have risen in almost exactly the same way as it actually did.

Indeed, (N#) can be established by very similar reasoning. The closest antecedent worlds where not all of the relevant neurons fire will be those worlds where all those neurons except for one or two fire the way they actually do. It is reasonable to suppose that a different activation of one or two neurons will still result in an arm rising that will be only slightly different from the actual one.

1.2. Counterexamples

There are many cases where counterfactuals with negations of conjunctions in their antecedents have truth-values that are at odds with those established on the standard analysis.

Example A. Suppose that Jane is allergic to cats. Her two closest friends, Alice and Bob, each get a cat. As a result, since Jane spends so much time with them, Jane suffers an allergic reaction. What would have happened if not both Alice and Bob had gotten a cat?

Suppose someone gives the following answer:

(A#) If Alice and Bob had not both gotten a cat, Jane would still have suffered an allergic reaction.

This statement seems to be false. The person making it as if forgets about the possibility allowed by the antecedent that neither Alice nor Bob could have gotten the cat, in which case *ceteris paribus* Jane would not have suffered an allergic reaction at all. What we can say is something weaker:

(A) If Alice and Bob had not both gotten a cat, Jane *might* have had an allergic reaction.

But it is false to say that she *would* have had it.

Yet ($A^{\#}$) seems to come out true on the standard analysis. *Ceteris paribus* the worlds where only one of Jane's friends does not get a cat are more similar to the actual world than the world where neither gets a cat. In those closest worlds, Jane will suffer the allergic reaction.

Example B. There is a box with one hundred 10g masses and a scale that can fit all of the masses (there are no other masses around). All of the masses from the box are actually placed on the scale, which indicates that their total mass is 1kg. What would have happened if not all of the masses had been put on the scale? Presumably, an exact answer depends on how many of the masses would have been put on the scale. *Ceteris paribus* the number can be any natural number between 0 and 99. So, if not all the masses had been put on the scale, it would have indicated between 0g and 990g (in increments of 10g).

The verdict licensed by the standard analysis is different. If one deems the number of the masses that are put on the scale as relevant to closeness then the closest worlds where not all the masses are put on the scale are those worlds where all but one or two masses are put on the scale. This means that in the closest possible worlds 99 or 98 (or close to 100) masses will be put on the scale, i.e. the following false counterfactual is established on the standard account:

($B^{\#}$) If not all masses had been put on the scale, the scale *would* have indicated 990g or 980g (or close to 1kg).

Intuitively, we have every reason to accept the weaker claim:

(B) If not all masses had been put on the scale, the scale *might* have indicated 990g or 980g (or close to 1kg).

but ($B^{\#}$) is too strong.

The standard account in effect ignores the richness of the logical structure of 'not all', just as it ignored the logical structure of 'not both' in the above example.

Example C. In general, it seems that the standard analysis will not allow enough attention to be paid to the logical structure of *not-all* antecedents. The pattern is well instantiated by the following two conditionals, which come out true on the standard semantics:

($1^{\#}$) If not everything were the same, everything would be pretty much the same.

($2^{\#}$) If not everything were different (if something were the same), everything would be the same.

($1^{\#}$) is true for the reasons we have seen above: from among the worlds that satisfy the antecedent 'not everything is the same', worlds where everything is almost the same are the closest to the actual world. ($2^{\#}$) is true because the worlds that satisfy the antecedent ('not everything is different' or equivalently 'something is the same') contain the actual world. The actual world is the closest world to itself. So what is the case in the actual world is what is the case in the closest antecedent worlds.[1]

Consider a concrete illustration of ($2^{\#}$), using our scale example. As before, we imagine that all of the masses from the box are actually placed on the scale, which indicates that their total mass is 1kg. What would have happened if at least one of the masses had been put on the scale? Presumably, an exact answer depends on how many of the masses would have been put on the scale. *Ceteris paribus* the number can be any natural number between 1 and 100. So, if at least one of the masses had been put on the scale, it would have indicated between 10g and 1000g (in increments of 10g).

[1] Indeed, while ($1^{\#}$) exemplifies the problem of disjunctive antecedents (§1.4), ($2^{\#}$) illustrates the combination of two problems: the problem of disjunctive antecedents and the problem of true antecedents.

The verdict licensed by the standard analysis is different. The class of the closest worlds where at least one of the masses is put on the scale includes the actual world where all masses are put on the scale. This means that we can establish the following false counterfactual on the standard account:

($C^{\#}$) If at least one of the masses had been put on the scale, the scale *would* have indicated 1kg.

Intuitively, we have every reason to accept the weaker claim:

(C) If at least one of the masses had been put on the scale, the scale *might* have indicated 1kg.

but ($C^{\#}$) is once again too strong.

As in the above cases, we can accept the weaker *might* counterfactuals but the *would* counterfactuals seem too strong. The standard analysis does not allow enough attention to be paid to the logical possibilities mentioned in the antecedent. In fact, one way to make the point is that the standard analysis seems to force the interpretation of a *not-all* antecedent as if it were an exclusive disjunction rather than an inclusive disjunction.[2]

1.3. The First Response to Lowe: Closeness and Similarity

It seems hard to deny[3] that worlds where all but one neuron fire the way they actually do, are the most similar to the actual world *if* all respects (including the behavior of the neurons) are

[2] On the standard analysis, if p and q are true in the actual world, the counterfactual $\sim(p \,\&\, q) \,\Box\!\!\rightarrow r$ will always have the same truth-values as $\sim(p \equiv q) \,\Box\!\!\rightarrow r$. Or equivalently: $(p \,\&\, q) \supset [(\sim p \vee \sim q) \,\Box\!\!\rightarrow r) \equiv ((\sim p \vee \sim q) \,\&\, \sim(\sim p \,\&\, \sim q)) \,\Box\!\!\rightarrow r)]$. In such cases, the full logical structure of the antecedent will never be taken into account.

[3] One might actually deny this by focussing on Lowe's insistence that *N* be construed as an event token. One could then use the standard analysis not just to show that if *N* had not occurred then the arm would have risen *almost like* it actually did, but rather to argue that if *N* had not occurred then the arm would have risen *exactly like* it actually did.

considered. One could have doubts, however, whether those worlds are the only most similar worlds to the actual world in all *relevant* respects.

In discussing the problem of true antecedents, Lewis suggests that 'perhaps our discriminations of similarity are rather coarse and some worlds different from [world] *i* are enough like *i* so that such small differences as there are fail to register' (Lewis 1973, p. 29). In other words, not all differences need to be taken to be relevant to the evaluation of similarity and closeness. In a later paper (Lewis 1979), Lewis provides a recipe for such evaluation, assigning different weights to various factors:

> (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. (p. 48)

It seems rather clear that the question of how many of the *k* neurons fire differently, how many of the 100 masses are placed on the scale, as well as how many of Jane's two friends not get a cat, belong to the fourth category. Lewis in fact admits that some cases (Tichý 1976; Jackson 1977) suggest that particular facts should *not* be relevant *at all*, though he does not fully endorse this suggestion.

One might very well argue that all of the above cases ($A^{\#}$), ($B^{\#}$) and ($C^{\#}$) – and many similar others – strongly suggest that the question of how many of negated conjuncts are satisfied is *irrelevant* to the evaluation of how close the antecedent worlds are to the actual world. If, among the closest possible worlds where at least one mass is placed on the scale, one admits

---

Let N^k-worlds be worlds where all those *k* neurons fire exactly in the same way as they actually do (so N^k-worlds are worlds, which are exactly like the actual world except that different tokens of the same type of neuron firings occur). N^k-worlds are closer to the actual world than any other worlds. The upshot of this point, however, ought to be that neuronal firings do not even determine the details of the arm movement. Moreover, this would set a dangerous precedent for it would seem that we would then have to claim that for any token *t* of type *T*, if *t* had not occurred then another token of the same type *T* would have occurred.

worlds where 1, . . . , 99, 100 masses are placed on the scale, we get the intuitively correct verdicts about ($C^{\#}$), which turns out to be false, as it should be. If, among the closest possible worlds where not all masses are placed on the scale, one admits worlds where 0, 1, . . . , 99 masses are placed on the scale, we get the intuitively correct verdict about ($B^{\#}$), which turns out to be false, as it should be. If among the closest possible worlds one admits worlds where neither Alice nor Bob get a cat, ($A^{\#}$) turns out to be false as well.

Likewise, one could argue, we should do the same when we are asking the question what would have happened if not all neurons had fired the way they actually did. One could argue that even N^0-worlds (worlds where none of the neurons $n_1, \ldots, n_k$ fire the way they actually did), not to mention N^1-worlds, . . . , N^{k-1}-worlds, N^k-worlds, are among the closest to the actual world. After all, all those worlds are governed by the same laws as the actual world, no miracles or other violations occur. Moreover, the subject's neuronal structure is the same. If the subject had more or fewer neurons, if they were differently connected or if they exhibited different efferent pathways, one might count such factors as contributing to a relevant dissimilarity. Such worlds, one might agree, would certainly not be the closest to the actual world. But the mere fact that the same neurons, hooked up in the same way, governed by the same laws of nature, fire differently than they actually do, should not be relevant to the question how close a given possible world is to the actual one.

If one accepts N^0-, . . . , N^k-worlds as the closest worlds then the neural counterfactual ($N^{\#}$) turns out to be false. Instead, two other counterfactuals are true:

> If *N* had not occurred, the arm might have risen.
>
> If *N* had not occurred, the arm might not have risen.

If *N* had not occurred and sufficiently many neurons had fired the way they actually did, the arm would have risen. If *N* had not occurred and sufficiently many neurons had fired differently than they actually did, the arm would not have risen.

Reflecting on this response to Lowe, one might claim that it is dialectically weak. After all, the exact determination of closeness is a highly controversial matter. There are two things to be said. First, this point can be turned around against Lowe. His argument for ($N^{\#}$) relies solely on his delimitation of closeness, which is a highly controversial matter. The above response is as weak as is Lowe's own argument. Second, we have identified a whole class of cases, which appear problematic for the standard account with Lowe's understanding of closeness.

1.4. The Second Response to Lowe: The Problem of Disjunctive Antecedents

Quite independently of how to determine closeness, another problem that has been brought up very early against the standard analysis of counterfactuals is relevant here, viz. the problem of disjunctive antecedents. It turns out that the standard analysis often gives unintuitive verdicts about counterfactuals with disjunctive antecedents. The problem is relevant here since, after all, the negation of a conjunction is equivalent to a disjunction of negations.

Many of the early critics of the standard analysis (Fine 1975; Ellis et al. 1977; Nute 1975; Creary & Hill 1975) have noted that it fails to accommodate our intuitions about counterfactuals with disjunctive antecedents. Consider the following example:

($3^{\#}$) If Spain had joined the Allies or the Axis, Hitler would have been pleased.

Spain was neutral during the Second World War but it was much closer to the Axis than to the Allies. Admittedly then worlds where Spain joins the Axis are closer than worlds where Spain joins the Allies. Thus, in the closest antecedent worlds, Hitler is pleased. Still a lot of people are

uncomfortable with ($3^{\#}$) and do not take it to be true because, they point out, if Spain had joined the Allies then Hitler would not have been pleased.

Many authors have suggested that the simplest way to explain this divergence of our intuitions is that we appear to accept the so-called principle of the simplification of disjunctive antecedents:

(SDA) $((p \vee q) \boxrightarrow r) \supset ((p \boxrightarrow r) \,\&\, (q \boxrightarrow r))$

If one accepts (SDA) then the argument for ($A^{\#}$), ($B^{\#}$), ($C^{\#}$), and ($N^{\#}$) is easily undercut. Consider the simplest allergy case. Given ($A^{\#}$) $\sim(A \,\&\, B) \boxrightarrow L$, which is equivalent to $[(\sim A \,\&\, B) \vee (A \,\&\, \sim B) \vee (\sim A \,\&\, \sim B)] \boxrightarrow L$, by (SDA), we obtain: $(\sim A \,\&\, \sim B)] \boxrightarrow L$, which is clearly false for the allergy case. If neither Alice nor Bob had gotten a cat, Jane would not have suffered an allergic reaction. So, one can argue, the reason why we take ($A^{\#}$) to be false is that we take one of its consequences (derived by means of (SDA)) to be false.[4]

The critics have pointed out, however, that the addition of (SDA) to possible-worlds semantics, while retaining an unrestricted principle of substitution of provable equivalents, leads i.a. to the loss of nonmonotonicity (Fine 1975; Ellis et al. 1977; Nute 1975; Loewer 1976). It suffices to substitute $p \vee (p \,\&\, q)$ for p in $p \boxrightarrow r$, to obtain $(p \,\&\, q) \boxrightarrow r$ by (SDA). Moreover, McKay and van Inwagen (1977) have identified counterexamples to (SDA) (for a recent defense of SDA, also from the counterexample, see Fine 2012a).

[4] It should be emphasized that an appropriate substitution needs to be made. Using just the de Morgan substitution of $\sim A \vee \sim B$ for $\sim(A \,\&\, B)$ will not lead to undesirable consequences. All we would then derive by (SDA) are the counterfactuals $\sim A \boxrightarrow L$ and $\sim B \boxrightarrow L$. However, these counterfactuals are quite plausibly true. If Alice had not gotten a cat then *ceteris paribus* Jane would have suffered an allergic reaction – we have no reason in this case to suppose that Bob would not have gotten a cat. The same holds for the other counterfactual. This is not a problem, however. All this shows is that not every logical consequence of ($A^{\#}$) is false.

Because the intuitions underlying (SDA) were quite powerful, various proposals have been put forward to try to do justice to them while holding off the unwelcome consequences (for a review, see Nute 1984a and Nute & Cross 2001). Two general substrategies of response have been pursued: to retain (SDA) while restricting the substitution principle (e.g. Nute 1975; Fine 2012b), or to retain the substitution principle and to reject (SDA) while preserving our intuitions about the cases by pursuing the translation lore, as Nute calls it. The adherents of the latter approach (e.g. Loewer 1976; McKay & van Inwagen 1977) suggested that the logical structure of some counterfactuals is in fact different from that suggested by the surface grammar: they are not counterfactuals with disjunctive antecedents rather they are conjunctions of two counterfactuals with the alleged disjuncts as antecedents.

On either of these approaches, the counterfactuals ($A^{\#}$), ($B^{\#}$), ($C^{\#}$), and ($N^{\#}$) may be argued to be false. On the substitution-restriction strategy, one would have to argue that the substitutions of $(\sim p \ \& \ q) \vee (p \ \& \sim q) \vee (\sim p \ \& \sim q)$ for $\sim(p \ \& \ q)$ are allowed.[5] On the translation-lore strategy, one would have to argue that counterfactuals such as ($A^{\#}$) ought to be understood as not having the form $(\sim p \vee \sim q) \boxright r$ but rather the form of a conjunction: $[(\sim p \ \& \ q) \boxright r] \ \& \ [(p \ \& \sim q) \boxright r] \ \& \ [(\sim p \ \& \sim q) \boxright r]$. To argue against the identity theorist, Lowe would have to show that either there is something wrong with the relevant substitutions or that it would be impossible to give the conjunctive reading to the counterfactuals.

Another type of approach involves an attempt to offer an alternative semantics for counterfactuals (Alonso-Ovalle 2009; Briggs 2012; Fine 2012b), whose verdict on the cases at

[5] Fine (2012b) develops a semantics, in which he distinguishes three concepts of implication. This enables him to restrict the substitution principle in a principled way. Once one restricts the substitution principle to Fine's strict equivalence, the substitution of $p \vee (p \ \& \ q)$ for p is blocked, thus blocking the argument for strengthening the antecedent. At the same time, $\sim(p \ \& \ q)$ is strictly equivalent both to $\sim p \vee \sim q$ as well as to $(\sim p \ \& \ q) \vee (p \ \& \sim q) \vee (\sim p \ \& \sim q)$, thus securing the intuitive verdict about the cases discussed in §1.2.

hand would be in line with our intuitions. What all of these approaches have in common is the insistence that all possibilities allowed by the antecedent be considered. They thus provide further reasons to think that the counterfactuals ($A^{\#}$), ($B^{\#}$), ($C^{\#}$), and ($N^{\#}$) are false. These reasons are not conclusive, they may be overridden, but they are there.

Whatever stance one takes in this debate, there is a methodological problem that arises from it for Lowe. Lowe's argument for ($N^{\#}$) relies *solely* on the standard possible-world semantics. The standard account has been questioned in part because it gives unintuitive results about counterfactuals with disjunctive antecedents, of which ($N^{\#}$) can be considered to be an instance.[6] In fact, the problem about disjunctive antecedents has sparked a search for ways to augment the account so as to prevent it from establishing such counterfactuals as true. Since Lowe's sole reason to accept ($N^{\#}$) is the appeal to the standard account, these are good reasons for us to think that Lowe has failed to give a satisfactory argument for the truth of the neural counterfactual.

His reliance on the standard semantics in this case is objectionable precisely because of its problems with cases of this sort. No matter how good the semantics is overall, the fact that it has led to the problem of disjunctive antecedents, which is far from resolved, should at least lead one to use it with caution in such cases. It may be perfectly rational to use it for other cases but it is objectionable to use it in cases where it has led to problems in the past and it is irrational to rely on it as the sole source of support for them.

[6] The problem has been explicitly raised for *not-both* counterfactuals by Nute (1984b).

1.5. The Truth about the Neural Counterfactual: 'not exactly all', 'almost all'

Lowe (1995) puts forward a principle, which might be used to see where some intuitive appeal for ($N^{\#}$) might come from. Lowe says:

> upon being presented with a counterfactual sentence, we naturally endeavour, if possible, to construe it according to a similarity-measure which will make it come out as true provided that we can think of such a measure which is not unduly strange. (p. 56)

Using this principle, we can identify some contexts where we would be willing to accept even the neural counterfactual.

Consider the claim that I could make seeing a physicist perform an experiment:

(4) If the force of 10.513 N had not been attached, the ball would have accelerated almost as it actually did.

The charitable physicist might agree to a statement like that taking me to mean that I consider only cases where a slightly smaller or greater force is attached. The statement might be better expressed with the addition of 'exactly'.[7]

($4^{@}$) If the force of *exactly* 10.513 N had not been attached, the ball would have accelerated almost as it actually did.

What this shows is that there is a way of charitably interpreting someone who makes such a claim as (4) by appropriately narrowing down the class of possible worlds, so that the consequent holds in them. There are a number of possible worlds where the force attached is not exactly 10.513 N but approximately that much, and where the ball would have accelerated in almost the way it actually did.

This is not say, however, that (4) means the same as ($4^{@}$). If the physicist was asked at a conference, say, what would have happened to the ball if the force of 10.513 N had not been

[7] This thought arose in correspondence with E.J. Lowe.

attached to it, in all likelihood she would give a response that would capture a much wider range of possibilities than does ($4^{@}$). Unless there are reasons having to do with the limitations of the experimental equipment or set up, she should be interpreting the question 'What would happen if the force of 10.513 N had not been attached?' as asking about all possible values of the force. Scientists are usually much more interested in what happens in general than in particulars.

We can put ourselves in a mindset where we would be answering the question about what would have happened in the cases under discussion in the way Lowe thinks we should. Consider the scale case. What would have happened if *not exactly all* masses had been put on the scale? In such a case, it might be appropriate to claim:

($B^{@}$) If *not exactly all* masses had been put on the scale, it would still have shown approximately 1kg.

Likewise, we can see ourselves agreeing with Lowe's answer to the following question: What would have happened if *not exactly all* neurons had fired the way they did?

($N^{@}$) If *not exactly all* neurons had fired the way they actually did, the arm would still have risen (albeit slightly differently).

The phrase 'not exactly all' presumably means 'almost all' or 'all with a couple of exceptions'. In such cases, and there is no problem with asserting such claims.

It should be clear, however, that neither ($B^{@}$) nor ($N^{@}$) are equivalent to ($B^{\#}$) and ($N^{\#}$), respectively. It is ($N^{\#}$) that Lowe needs in order to establish the claim that the non-occurrence of the complex neural event token would still have led to a similar arm movement. It is ($N^{\#}$) that gives Lowe the basis for asserting that counterfactual properties of the two token events are different.

1.6. Conclusion

Let us take stock reminding ourselves of the dialectical setting. To argue against the token-identity thesis, Lowe needs to establish the truth of the neural counterfactual. I have argued that he has *failed* to do so. There are clearly false counterfactuals, whose truth can be established using Lowe's form of argument for the neural counterfactual (§1.2). The weakness of the argument is exposed by reflecting on the fact that it crucially trades on two issues that have been controversial for the standard possible world semantics: the notion of closeness (§1.3) and the problem of disjunctive antecedents (§1.4). I have argued that the defender of identity theory can find ample room to argue that the neural counterfactual is in fact false. I have further identified some contexts where a counterfactual similar to the neural counterfactual is true (§1.5) but I argued that this does not suffice to uphold Lowe's argument.

These results are more than sufficient to undercut Lowe's argument against the token-identity thesis. I will proceed to show that one may very well doubt that Lowe has established the truth of the mental counterfactual.

2. The Argument for the Mental Counterfactual

Lowe's argument for the mental counterfactual is given in the following quote:

> if *D* had not occurred – if the agent had not made the very act of choice that he or she did to raise the arm – then the arm *would not have risen at all*. It is, I suggest, quite incredible to suppose that if the agent had not made *that* very decision, *D*, he or she would have made a decision virtually indistinguishable from *D* – in other words, *another* decision to raise the arm in the same, or virtually the same, way. On the contrary, if the agent had not made *that* decision, then he or she would either have made quite a different decision or else no decision at all. Either way – assuming that there is nothing defective in the agent's nervous system – the arm *would not* have risen almost exactly as it did. (p. 105, original emphases)

The argument appears to have the following structure:

(M1) It is false that: if *D* had not occurred at *t* then another *D*-like event would have occurred at *t*.

(M2) If *D* had not occurred at *t* then no other *D*-like event would have occurred at *t* (the agent 'would either have made quite a different decision or else no decision at all', p. 106)

(M3) If no *D*-like event had occurred at *t* then no *B*-like event would have occurred at *t*′.

(M4) If *D* had not occurred at *t*, then no *B*-like event would have occurred at *t*′ ('If *D* had not occurred, my arm would not have risen at all', p. 105) (from (M3), (M2))

The conclusion (M4) follows from (M2) and (M3). Presumably, (M3) restates the commonsensical belief that our decisions (to raise an arm) are causally efficacious – that they are causally responsible for arm risings. I am not going to challenge (M3). (M1) is supported by Lowe's argument from the fine-grainedness of mental contents, to which I will turn in a moment. I will accept (M1) for the purposes of the argument.

The crucial step in the argument, and one which I will challenge, is (M2): if a given token of my decision to raise my arm had not occurred at *t* then I would not have made, at *t*, another decision to raise my arm. It is not clear from the text how (M2) is justified. (M2) does not follow from (M1) – at least not on Lewis' possible-world semantics, on which Lowe relies in the paper.[8] The only argument Lowe offers is the argument from fine-grainedness of content. We will see, however, that, on one reading, the fine-grainedness argument supports (M1) but not (M2) (§2.1), while on another (§2.3), it presupposes, among others, the problematic neural counterfactual. I argue that Lowe has not established that (M2) is true – it can be plausibly argued to be false as a

[8] One way to argue for (M2) would be via accepting the conditional excluded middle. The question whether the principle of excluded middle holds for counterfactuals has famously divided the founders of the standard approach: Lewis (1973) rejects it while Stalnaker (1968; 1981) accepts it. The principle is not a theorem on Lowe's own semantics (1983; 1995).

general principle (§2.2) and in fact one may have reasonable doubts whether it is ever true (§2.4).

2.1. The Fine-Grainedness Argument

Lowe argues that mental contents are incapable of individuating decisions in a way that is fine-grained enough to match the individuation conditions of physical movements. Arm risings can be individuated by appeal to their precise trajectories, for example, but it would be hard to suppose that decisions-to-raise-arm can match arm risings in this respect.[9]

Let us agree with Lowe on (M1): there will be many more different possible arm-rising tokens than decision-to-raise-arm tokens. So the view that Lowe challenges, viz.

(M1*) If *D* had not occurred at *t* then another *D*-like event *would* have occurred at *t*.

is indeed quite plausibly false. Lowe in effect argues that someone who wanted to argue for (M1*) would have to secure ways of individuating mental contents to match the plethora of ways available for individuating bodily movements, which seems implausible.

If Lowe's argument from fine-grainedness of mental contents is construed in this way (we will see another way to construe it in §2.3), it justifies (M1) but not (M2). From the fact that there are many more different possible arm-rising tokens than decision-to-raise-arm tokens, it follows at most that given that a particular decision-to-raise-arm token *D* does not occur, it is not the case that *in all* relevant worlds another decision-to-raise-arm token will occur because there will not be enough decision-to-raise-arm tokens to match the possible arm-rising tokens. It does *not* follow that *there is no* relevant world where another decision-to-raise-arm token will occur

[9] One could challenge this claim by appealing to *de re* intentions (see e.g. Wilson 1989). This challenge is potentially serious to Lowe because Wilson's conception is offered in complete abstraction from the debate about identity theory, so one could not argue that it is question-begging.

(here the fact that the cardinality of one of the sets outruns the other is irrelevant). From the fact that there is not enough food to feed everyone it does not follow that nobody can be fed. So construed the fine-grainedness argument supports (M1) but not (M2).

2.2. A Plausibly False Premise

Even if one agrees with Lowe on (M1), it is unclear why (M2) ought to be accepted. Lowe's opponent might well agree with him that it is wrong to think that (M1*) is true, i.e. that if *D* had not occurred then another *D*-like event *would* have occurred. But she might still accept the weaker claim that:

(M2*) If *D* had not occurred at *t* then another *D*-like event *might* have occurred at *t*.

(M2*) is inconsistent with (M2). So in order to establish (M2), Lowe would have to show that (M2*) is false. The problem is that (M2*) is quite often plausibly true, which makes (M2) plausibly false.

To consider whether (M2) is true, we must bear in mind the dialectical setting. Lowe claims that:

> Contentful mental acts such as decisions are [. . .] *individuated* at least partly by their contents – and yet their contents surely cannot be as fine-grained as the physicalist's conjectured contention would appear to demand. How, exactly, would the *content* of the decision that, supposedly, would have occurred if *D* had not occurred have differed from the content of *D*? (p. 106, original emphases)

Lowe is concerned here that the physicalist not presuppose the identity thesis in the argument. So he would reject as question-begging the suggestion that we individuate mental states not in terms

of contents but in terms of the physical characteristics of the physical states that are allegedly identical with the mental states.[10]

Lowe writes as if his challenger could appeal *only to contents* to individuate decisions. But Lowe's opponent (physicalist or not) may invoke at least three ways of individuating decision tokens: (a) in terms of the manner in which decisions are taken by the agent, (b) in terms of the deliberation process, which results in the decision, and, of course, (c) in terms of their content. One can plausibly argue for the truth of (M2*) using *any* one of these possible ways of individuating decision tokens. Let's consider these ways in turn.

(a) Decisions can be made in different ways: the agent might have hastily decided to raise her arm or she might have confidently decided to raise her arm. The decision tokens would differ (one would be made in haste, the other – confidently) but their content would be the same, viz. to raise the arm. So, it is plausible to argue that if *D* had not occurred, the agent might have taken another decision token of the same type in a somewhat different way – in a less hasty or in a more confident manner, for example.

(b) Decisions are often the results of deliberation. It is quite intelligible to suppose that a different decision token of the same type to raise an arm is reached via a slightly different process of deliberation. Again, it is plausibly argued that if *D* (which was, as a matter of fact, reached through deliberation δ) had not occurred, the agent might have engaged in a different

[10] The question whether the identity theorist may appeal to individuation conditions that rely on the identity thesis is actually not settled as straightforwardly. Much will depend on what the import of Lowe's argument is supposed to be. Consider two possibilities. First, if we were trying to find out which view (identity theory or dualism) follow from our ordinary commitments couched in the sort of counterfactuals we are willing to accept, then one might indeed think that in such a context it would be question-beginning for the identity theorist to 'sneak in' the identity thesis. However, if Lowe simply means to argue 'against psychoneural identity theories' (p. 103) and to conclude that 'the decision *D* cannot be identical with the neural event *N*' (p. 107), then in such a context the identity theorist is clearly in her right to defend herself by any means that her theory can offer. This point by itself shows that Lowe's conclusion is far too strong. Of course, her defence will be stronger if she does not need to appeal to the identity thesis.

deliberation δ′, which would have resulted in a different decision token *D′* of the same type as *D*, i.e. to raise the arm. In fact, one does not need to suppose that the deliberation was dramatically different – the agent may have considered the same reasons but proceeded in a different order, for example.

(c) Let us finally consider whether a decision with a different content *could* lead to the same type of movement (arm rising). *Prima facie* there is not much problem in coming up with an example where a decision with a different content leads to a similar type of movement. I will consider two classes of examples.

It would be sufficient to imagine that the agent have an intention with a more specific content, e.g. to raise the arm quickly or to raise the arm vertically rather than to raise the arm (in just any way). So if *D* (the agent's decision token *to raise the arm*) had not occurred, the agent *might* have decided *to raise the arm quickly* (*D″* might have occurred). The ensuing arm rising movement could have been very similar to the actual one.

One may think that this suggestion is subject to the following objection: if the agent raises her arm quickly, she still raises her arm; so if the agent does not decide to raise an arm, she cannot decide to raise her arm quickly. But the objection fails for two reasons. First, it is of course true that if φing implies ψing then if α φs then α ψs. If the agent raises her arm quickly then she raises her arm. But it is rather widely acknowledged that intentions are not closed under implication, not even under known implication. If an agent intends to win a competition, it does not follow that she intends either to win or to lose the competition. It is possible for someone to intend to raise her arm in a specific way without intending to raise her arm (in just any way). Think of an actress who intends to raise her arm so as to express despair. She might very well not

intend to raise her arm *period*. In such a case, her intention would *not* be realized if she had just flipped her arm upward.[11]

Second, the objection seems to confuse types with tokens. Even if one agreed that to decide to raise an arm slowly is to decide to raise an arm, this is a point about decision *types*. One could still hold (M2*) to be true. One would then say: if *D* (decision token of the type *to raise an arm*) had not occurred, there might have occurred a different decision token *D‴* (of the type *to raise an arm slowly*), which is also of the type *to raise an arm*. Just because *D‴* is of the same type as *D* does not show that *D‴* is the same token as *D*.

Greater specificity is not the only way in which contents can differ, however. It would be possible for the agent to decide to stretch, in which case he *might* also raise his arm in a way that might be indistinguishable from the way he would have raised it had he decided to raise an arm. One might object here that if the agent decided to stretch rather than raise the arm (to ask a question, say), the agent would have moved much more differently for very probably the agent would not have wanted to give the impression of intending to ask a question and so would have stretched the arm in such a way as to prevent a possible misinterpretation of his movement. The objector is right to point out that if the agent is fully aware of the situation, of the consequences of his actions, of the possible interpretations of his actions by others, then he *might* do everything in his power to perform the action in a way to prevent possible confusion. But again it would be too strong to claim that the agent *would* do everything in his power to prevent possible confusion. This may very well depend on other factors. Suppose that Alexia decides to raise her

[11] One might worry here that I am assuming that the agent may only have one intention. But this is not so. It is, of course, possible for someone to have both the intention to φ and the intention to φ or to ψ. An actress may intend to raise her arm to express despair and she may have an intention to raise the arm in despair or to raise it in some other way. My point is only that it is also possible for someone to have only one intention. The mere fact that one has an intention to φ does not mean that one thereby also has an intention to φ or to ψ.

arm to ask a question in order to call attention to herself. Now it may very well be that if Alexia had not taken the decision to raise her arm to ask a question, she might have decided to stretch or to laugh loudly in order to call attention to herself. In such case, if Alexia had decided to stretch rather than raise her arm, it is unlikely that she would very much care about a possible misinterpretation of her movement. Indeed, if Alexia had not decided to raise her arm, her arm might have moved almost as it actually did or it might not have moved at all.

Lowe has failed to exclude the possibility that (M2*) is true, i.e. that if the agent had not taken the decision to raise her arm, she might have still taken another decision to raise her arm. I have shown that one can plausibly argue for (M2*) by appealing either to the manner in which the decision is taken or to the deliberation process of which the decision is the result or indeed to the content of the decision.

2.3. The Fine-Grainedness Argument Again

Here is another way of employing the observation about the differences in the grain of the individuation conditions. Consider the following passage from Lowe:

> How, exactly, would the *content* of the decision that, supposedly, would have occurred if *D* had not occurred, have differed from the content of *D*? If the putative difference in their contents is to match the *very slight* difference between the bodily events that are supposed to ensue from them, then a degree of fine-grainedness must be attributed to those contents that . . . is utterly implausible from a psychological point of view. (p. 106, original emphases)

Lowe seems to argue here that if the decision token *D* (which is allegedly identical to *N*) had not occurred, then – as we know from the argument for the neural counterfactual – another neural event token *N** (very similar to *N*) would have occurred, which would have produced an arm movement *B** (very similar to the actual movement *B*). The identity theorist would think that *N** is identical to a putative decision token *D**. So, *D** would have to have produced the arm

movement *B**. But the difference between *B* and *B** is miniscule – in fact the arm movement tokens might be indistinguishable to us. Lowe claims that there is no psychologically accessible decision token *D**, which would differ from *D* in a way that would match the small difference between *B** and *B*. So, he argues that it is implausible to suppose that *under these conditions* another decision token to raise an arm would have occurred.

It should be stressed that this argument can be read as being directed *also* against (M2*). Lowe never explicitly considers the possibility of (M2*) but the above rendition of the argument suggests that under the counterfactual conditions laid out by the neural counterfactual Lowe would have a reason for thinking that it is *not* the case that there even *might* have occurred a decision token that would have led to a similar bodily movement.

The argument is highly problematic, however, for at least two reasons. First, it crucially depends on the neural counterfactual, whose truth we have good reasons to doubt (§1). If Lowe's argument for the mental counterfactual depends on his argument for neural counterfactual then my case against Lowe's argument for the neural counterfactual is *already* a case against Lowe's argument for the mental counterfactual. Second, the above argument seems to rely on a matching principle, which seems rather implausible and in any case Lowe does not justify it. Why should we believe that bodily movements that differ from each other in a small way would have to be produced by decisions that differ from each other in a correspondingly small way? The principle might perhaps be held by someone who thought that decisions determine the precise parameters of a bodily movement. Such a view is very implausible, however, and Lowe explicitly rejects it. It is unclear then why Lowe seems to demand that a miniscule difference in the resulting bodily movements be matched by a similarly miniscule difference in the decision contents. It seems perfectly plausible to suppose that a very similar bodily motion *might* have been produced if the

agent had decided to raise an arm for its own sake as well as if she had decided to raise her arm to ask a question.

2.4. Is the Mental Counterfactual Ever True?

What I have shown is that Lowe's arguments do not establish that the mental counterfactual holds in general. While it seems plausible to suppose that if the mental event token had not occurred then the arm movement *might* have *not* occurred at all, it is also plausible to suppose that if the decision token had not occurred then a similar arm movement *might* have occurred. To argue for the mental counterfactual Lowe would have to provide a general argument to the effect that (M2) is true, i.e. that if a particular decision token had not occurred then a similar decision token would not have occurred. Lowe's fine-grainedness argument does not establish that (M2) is true (§2.1, §2.3). We have seen several plausible ways to argue that there are occasions where (M2) is false and its contrary (M2*) is true (§2.2).

So far I have not argued that (M2*) is true in general. It should be pointed out, however, that my arguments against Lowe's mental counterfactual do not depend on establishing that (M2*) holds in general. Lowe takes it that for *any* purportedly identical neural event token and decision token, he has general arguments, which show that the non-occurrence of the neural event token leads to a certain type of event while the non-occurrence of the decision token leads to the non-occurrence of that type of event. I have argued against the mental counterfactual by suggesting that there are many cases where it does not hold, where (M2*) rather than (M2) is true.

One might think that this opens the possibility that there might be cases, for which Lowe's mental counterfactual is true.[12] One might argue in particular that it may very well be that if Mary had not decided to raise an arm on a particular occasion, her arm would not have moved in the way it actually did, for just the reasons that Lowe suggests – if she had not decided to raise an arm, she would have decided to do something entirely different or would not have taken any decision. If one grants that, however, then I would insist that there are *also* cases where the contrary of Lowe's counterfactual is true. It may be that if Denise had not taken the decision to raise her arm (e.g. as a result of deliberating in the way she did), she would have still decided to raise her arm (as a result of a slightly different deliberation), in which case her arm would still have risen. Finally, there are cases where the *would*-counterfactuals are both false. It may be that if Alexia had not taken the decision to raise her arm in order to call attention to herself, she might have decided to stretch in order to call attention to herself or she might have decided to laugh loudly in order to call attention to herself. (I consider the implications of accepting such counterfactuals for the whole argument in §3).

I must admit, however, that I do not wholeheartedly embrace this possibility and find (M2*) to be most plausible in general. The counterfactuals about Mary's, Denise's, and Alexia's decisions to raise an arm are plausible as long as they are construed as counterfactuals about *types* of mental events. They are implausible if one construes the decisions as tokens, as Lowe would have to. Consider the counterfactual about Mary as an example.

We may grant that if Mary had not decided to raise an arm (i.e. if no event token of the type *Mary decides to raise an arm* occurred), her arm would not have moved in the way it actually did – if she had not decided to raise an arm, she would have decided to do something

[12] I would like to thank the anonymous reviewer for this journal for suggesting this line of response.

entirely different or would not have taken any decision. However, what reason would we have for thinking that if a particular event token of the type *Mary decides to raise an arm* had not occurred, then another event token of the type *Mary decides to raise an arm* would not have occurred? Note that it is perfectly consistent to hold the mental counterfactual about event types while refusing that it holds for event tokens. One may very well think that if no event token of the type *Mary decides to raise an arm* had occurred then Mary would have decided to do something different or would not have taken any decision. But it does not follow from this that if a particular event token of the type *Mary decides to raise an arm* had not occurred then Mary would have decided to do something different or would not have taken any decision. *Prima facie* it seems that she *might* have taken another decision token to raise an arm. She might have done so in a somewhat different way, or a moment later, or as a result of a slightly different deliberation, or the decision may have had a slightly different content. It seems bizarre to think that Mary *would not* have taken another decision to raise an arm unless we are speaking about decision types rather than decision tokens. Pending reasons to the contrary, I think that (M2*) is plausible in general.

In any event, whether (M2*) is plausible for all decision tokens or only for some, (M2) is false as a general claim. This is enough to undercut Lowe's general argument for the mental counterfactual.

3. On the Counterfactual Implications of the Token-Identity Thesis

I have argued that Lowe's argument from counterfactual implications against the token-identity thesis fails. While there are reasons to believe that the neural counterfactual is false, at the very

least my argument shows that Lowe has failed to establish its truth. Lowe's argument for the mental counterfactual relies, as I argued, on a premise that is plausibly false. Once again there are reasons to believe that the mental counterfactual is false, but at the very least we have seen that Lowe has failed to establish its truth.

In all this argumentative turmoil, it might be useful to point out that there is a perfect match between the following 'base' counterfactuals:

> If *N* (the neural event token) had not occurred then *B* (the particular arm rising token) would not have occurred.
>
> If *D* (the particular arm raising decision token) had not occurred then *B* (the particular arm rising token) would not have occurred.

Can we say anything about the sort of counterfactuals that Lowe is interested in: whether a *B*-like event would or would not have occurred? I have argued that Lowe's arguments do not establish either the truth of the neural counterfactual or the truth of the mental counterfactual. Given just the initial description of the case and Lowe's arguments, the weaker *might*-counterfactuals seem much more plausible on either the mental or the neural side.

> If *N* had not occurred, a *B*-like event might have occurred (the arm might have risen).
>
> If *N* had not occurred, a *B*-like event might not have occurred (the arm might not have risen).
>
> If *D* had not occurred, a *B*-like event might not have occurred (the arm might not have risen).
>
> If *D* had not occurred, a *B*-like event might have occurred (the arm might have risen).

A physicalist can feel satisfied by the consistency between the token-identity thesis and these counterfactuals. Appearances to the contrary, Lowe's argument does not establish that the consistency is threatened.

In §2.4, I have considered the possibility that one might argue that even if the mental counterfactual is not true in general, it might be true in some cases. I have argued against this suggestion but suppose that one could find some reasons to challenge my arguments. Suppose that it could be established that the counterfactuals about Mary's, Denise's and Alexia's decisions hold not just for decision types but also for decision tokens. Would this suggestion not reintroduce an asymmetry between the kinds of counterfactuals one might hold on the mental and the neural side? The matter would certainly have to be investigated further. *Prima facie* one might think that the physicalist may also argue that, on various occasions, the background causal conditions may also settle it that the arm not only might have risen but that it would have risen, or they may settle it that the arm would not have risen or they may not settle it whether the arm would or would not have risen. In other words, the physicalist might argue that similar counterfactuals obtain on the neural side. Of course, this by itself would not establish the consistency of the mental and neural counterfactuals for the question whether they are true in the same cases could still be raised. So there might still be room for an argument against identity theory in the vicinity. However, this would have to be quite a different argument than the one offered by Lowe. My point has been that Lowe's argument against the identity theory is unsuccessful. Indeed, I doubt that an appeal to the semantics of counterfactuals can settle these matters.[13]

[13] I would like to thank an anonymous reviewer for this journal for very helpful comments on a previous version of the paper. I am sincerely grateful to E.J. Lowe, who was extremely generous with his time in discussing some of the issues, as well as to Mariusz Grygianiec and Joanna Odrowąż-Sypniewska for their comments and discussions.

Institute of Philosophy

University of Warsaw

Bibliography

Alonso-Ovalle, L. (2009). 'Counterfactuals, Correlatives, and Disjunction', *Linguistics and Philosophy* 32, pp. 207-244.

Briggs, R. (2012). 'Interventionist Counterfactuals', *Philosophical Studies* 160, pp. 139-166.

Creary, L. G., & Hill, C. S. (1975). 'Review of D. Lewis' Counterfactuals', *Philosophy of Science* 42, pp. 341-344.

Ellis, B., Jackson, F., & Pargetter, R. (1977). 'An Objection to Possible-World Semantics for Counterfactual Logics', *Journal of Philosophical Logic* 6, pp. 355-357.

Fine, K. (1975). 'Critical Notice of D. Lewis' Counterfactuals', *Mind* 84, pp. 451-458.

Fine, K. (2012a). 'A Difficulty for the Possible Worlds Analysis of Counterfactuals', *Synthese* 189(1), pp. 29-57.

Fine, K. (2012b) 'Counterfactuals without Possible Worlds', *The Journal of Philosophy* 109, pp. 221-246.

Gillies, T. (2007). 'Counterfactual Scorekeeping', *Linguistics and Philosophy* 30, pp. 329-360.

Jackson, F. (1977). 'A Causal Theory of Counterfactuals', *Australasian Journal of Philosophy* 55, pp. 3-21.

Lewis, D. (1973). *Counterfactuals.* Cambridge, MA: Harvard University Press.

Lewis, D. (1979). 'Counterfactual Dependence and Time's Arrow', *Nous* 13, pp. 455-476.

Loewer, B. (1976). 'Counterfactuals with Disjunctive Antecedents', *Journal of Philosophy* 73, pp. 531-537.

Lowe, E. (1983). 'A Simplification of the Logic of Conditionals', *Notre Dame Journal of Formal Logic* 24, pp. 357-366.

Lowe, E. (1995). 'The Truth about Counterfactuals', *The Philosophical Quarterly* 45, pp. 41-59.

Lowe, E. (2006). 'Non-Cartesian Substance Dualism and the Problem of Mental Causation', *Erkenntnis* 65, pp. 5-23.

Lowe, E. (2008). *Personal Agency. The Metaphysics of Mind and Action.* Oxford: Oxford University Press.

McKay, T., & van Inwagen, P. (1977). 'Counterfactuals with Disjunctive Antecedents', *Philosophical Studies* 31, pp. 353-356.

Earlier version of my argument has been published in Polish in *Filozofia Nauki* 21-1 (2013): 91-112. The work on this paper has been made possible in part by an NCN grant (DEC-2012/05/B/HS1/02949).

Nute, D. (1975). 'Counterfactuals and the Similarity of Worlds', *Journal of Philosophy* 72, pp. 773-778.

Nute, D. (1984). 'Conditional Logic', in D. Gabbay & F. Guenthner (eds.), *Handbook of Philosophical Logic*, vol. 2. Dordrecht: Reidel, pp. 387-439.

Nute, D. (1984). *Topics in Conditional Logic.* Dordrecht: Reidel.

Nute, D., & Cross, C. B. (2001). 'Conditional Logic', in D. Gabbay & F. Guenthner (eds.), *Handbook of Philosophical Logic*, 2nd edition, vol. 4. Dordrecht: Kluwer, pp. 1-98.

Stalnaker, R. (1968). 'A Theory of Conditionals', in *Studies in Logical Theory. American Philosophical Quarterly Monograph Series* 2. Oxford: Blackwell, pp. 98-112.

Stalnaker, R. (1981). 'A Defense of Conditional Excluded Middle', in W. Harper, R. Stalnaker & G. Pearce (eds.), *Ifs: Conditionals, Belief, Decision, Chance, and Time*. Dordrecht: Reidel, pp. 87-104.

Tichy, P. (1976). 'A Counterexample to the Stalnaker-Lewis Analysis of Counterfactuals', *Philosophical Studies* 29(4), pp. 271-273.

Wilson, G. (1989). *The Intentionality of Human Action.* Stanford: Stanford University Press.

Appendix: The Neural Counterfactual on Lowe's Semantics

While Lowe uses Lewis's semantics to make his argument, he has his own preferred approach to counterfactuals (see Lowe 1983; 1995), in which he defines a counterfactual in terms of a strict conditional in the following way:

(Def) $(p \mathrel{\Box\!\!\rightarrow} r) =_{\text{def}} \Box(p \supset r) \,\&\, (\Diamond p \vee \Box r)$

The principle of simplification of disjunctive antecedents does not hold on Lowe's semantics because $\Diamond(p \vee q) \supset \Diamond p$ is not a theorem. However, a weaker principle

(SDA′) $((p \vee q) \mathrel{\Box\!\!\rightarrow} r) \;\&\; \Diamond p \supset (p \mathrel{\Box\!\!\rightarrow} r)$

is a theorem on Lowe's account.[14]

Let me provide a pattern of argument against Lowe's neural counterfactual taking ($A^{\#}$) as a model. One begins by making sure that we understand what would happen in various counterfactual situations:

(1) $\sim A \mathrel{\Box\!\!\rightarrow} J$ (If Alice had not gotten a cat, Jane would have suffered from the allergy; if only Bob had not gotten a cat, Alice would still have got it, and Jane would develop the symptoms.)

(2) $\sim B \mathrel{\Box\!\!\rightarrow} J$ (analogously)

(3) $(\sim A \,\&\, \sim B) \mathrel{\Box\!\!\rightarrow} \sim J$ (If neither Alice nor Bob had gotten a cat, however, Jane would not have suffered from allergy.

It should be stressed that (1)-(3) do not seem to be controversial at all. From (3), (Def) and the plausible assumption that it is not necessary that Jane suffers from allergy, we derive:

(4) $\Diamond(\sim A \,\&\, \sim B)$

Suppose then that someone asserts:

($A^{\#}$) $\sim(A \,\&\, B) \mathrel{\Box\!\!\rightarrow} J$

Substituting an equivalent proposition in the antecedent of ($A^{\#}$), we obtain:

($A^{\#\#}$) $[(\sim A \,\&\, B) \vee (B \,\&\, \sim A) \vee (\sim A \,\&\, \sim B)] \mathrel{\Box\!\!\rightarrow} J$

Using (SDA′) as well as the possibility claim (4), we obtain the contradictory of (3):

($3^{\#}$) $(\sim A \,\&\, \sim B) \mathrel{\Box\!\!\rightarrow} J$

[14] Given (1) $((p \vee q) \mathrel{\Box\!\!\rightarrow} r)$ and (2) $\Diamond p$, by (Def) we obtain (3) $\Box((p \vee q) \supset r)$ and (4) $\Diamond(p \vee q) \vee \Box r$. We can derive (5) $\Box(p \supset r)$ from (3) (for an arbitrary world, suppose that p, so also $p \vee q$, and by (3) r). And we can also derive (6) $\Diamond p \vee \Box r$ (by addition to (2)). By (Def), (5) and (6) we obtain the conclusion: $p \mathrel{\Box\!\!\rightarrow} r$.

Since ($3^{\#}$) is false (if neither Alice nor Bob had gotten a cat, Jane would not have suffered the allergic reaction), it follows that either ($A^{\#}$) is false or (4) is false. But (4) is true since (3) is true, so ($A^{\#}$) must be false.

It might seem that the result is obtained by means of additional premises (like (1)-(3)). However, it is arguable that those additional premises are really part of the background understanding of what happens. Consider the scale case. As we are explaining the case, we assume naturally that the scale is reliable, which means *inter alia* that:

(r) If n masses were placed on the scale, it would show $10n$ g (for $n = 0, \ldots, 100$)

By (Def) we derive the claims that either it is possible that n masses are placed on the scale or it is necessary that the scale shows $10n$ g. The second disjunct drops out since it is not necessary that the scale show a particular reading, leaving us with:

(n) It is possible that n masses are placed on the scale (for $n = 0, \ldots, 100$).

If so, then we can substitute the antecedent of ($B^{\#}$) with an equivalent one to obtain:

($B^{\#\#}$) If either no masses had been put on the scale, or only 1 mass had been put on the scale, or only 2 masses had been put on the scale, or . . . , or 99 masses had been put on the scale, the scale *would* have indicated 990g or 980g (or close to 1kg).

From ($B^{\#\#}$) and (n), using (SDA′) we can obtain, for example:

If only 1 mass had been put on the scale, the scale would have indicated 990 g or 980g (or close to 1kg).

which is clearly false – is in fact in contradiction with (r).

Similarly for the neural counterfactual we can argue that it is part and parcel of the background understanding of how the neural system of the agent works that if only one (or few) of the k neurons had fired in a certain way (differently than they actually did), the arm would still

have risen, but if sufficiently many (say, at least *j*) neurons had fired differently, the arm would not have risen. To accept such counterfactuals as true is, given Lowe's semantics, in effect to accept that it is possible for sufficiently many neurons to fire differently than they did. This then means that we can use (SDA′) to derive from the neural counterfactual the unwelcome consequence that if sufficiently many (at least *j*) neurons had fired differently, the arm would still have risen (in contradiction with the original claim). We thus have reasons to think that the neural counterfactual is false even on Lowe's own semantics.

Lowe would no doubt take the above as an argument that we should abandon the possibility claim, i.e. deny that it was possible for sufficiently many (at least *j*) to have fired differently and, consequently, we should consider the claim that if sufficiently many (at least *j*) neurons had fired differently, the arm would not have risen to be false. In other words, because Lowe accepts that

($N^{\#}$) if not all neurons had fired the way they did, the arm would have still risen

he denies that the following claims are true:

> If all neurons had fired differently than they did, the arm would not have risen.
>
> . . .
>
> If *j* neurons had fired than they did, the arm would not have risen.

It is his denial of the above claims, which will constitute a barrier to a common understanding of the case between Lowe and his opponents. I doubt that there is anybody who *ceteris paribus* would be willing to deny the truth of such claims.