Proceedings of SALT 34: 000–000, 2024

New Experimental Evidence Against the Similarity Approach to Conditionals[*]

Dean McHugh
University of Amsterdam

Tomasz Klochowicz
University of Amsterdam

Abstract The similarity approach to conditionals (Stalnaker 1968; Lewis 1973) predicts Reciprocity to be valid: whenever $A > B$, $B > A$ and $A > C$ are true, $B > C$ is true too (where $A > B$ denotes *if A would B*). We ran an experiment to test the validity of this rule. Strikingly, half of our participants judged the rule invalid, i.e. judged in at least one scenario that it does not preserve truth. Our data also challenge Kratzer's (2012) and Fine's (2012) semantics of conditionals, but we show that McHugh's (2022; 2023) aboutness approach can account for our data.

Keywords: counterfactuals, conditionals, reciprocity, CSO

1 Introduction

For the last half century, semanticists have been attracted to the idea that a *would*-conditional "if A, would C" (henceforth denoted $A > C$) is true just in case C is true at the most similar world(s) to the actual world where A is true. This is arguably the dominant approach to the semantics of *would*-conditionals, proposed by Stalnaker (1968) and Lewis (1973), influentially extended to modality in general via Kratzer's (1981a) ordering source, and taught in numerous textbooks and course syllabi today.

The similarity approach validates the following rule, known as *Reciprocity*.[1]

Reciprocity. Whenever $A > B$, $B > A$, and $A > C$ are true, $B > C$ is true.

Reciprocity says that if A and B conditionally imply each other, then they are intersubstitutable salva veritate in conditional antecedents. This rule is part of the logic of similarity approach to conditionals: no matter how one intuitively interprets the similarity order, Lewis's and Stalnaker's semantics predict it to be valid.[2]

[*] We thank anonymous SALT reviewers for insightful feedback. This work is supported by *Nothing is Logical* (NihiL), an NWO OC project (grant no 406.21.CTW.023). For discussion we are grateful to Tanya Bondarenko, Paul Egré, and the audience of the Linguae Seminar at Institut Jean-Nicod.

[1] The name Reciprocity comes from Egré & Rott (2021). Nute (1980) calls it 'CSO', and Starr (2022) calls it 'Substitution of Subjunctive Equivalents'.

[2] By validity here we mean the usual notion: for every model and world, if $A > B$, $B > A$, and $A > C$ are all true in that model at that world, then $B > C$ is too.

We ran an experiment to test the validity of Reciprocity. In three separate scenarios we asked participants to judge sentences that form an instance of the rule. Strikingly, half of the participants judged the rule invalid, i.e. judged in at least one scenario that $A > B$, $B > A$, and $A > C$ are all true but $B > C$ is not. On average, in 28% of cases where a participant accepted the premises they rejected the conclusion.

The paper proceeds as follows. Section 2 shows that Lewis's and Stalnaker's semantics validate Reciprocity. In Section 3 we discuss previous alleged counterexamples to Reciprocity, all of which have a different structure than the counterexamples we tested. Section 4 presents our experimental design and results. In Section 5 we show that our results cannot be accounted for by the semantics of conditionals proposed by Kratzer (2012), Fine (2012), and Ciardelli, Zhang & Champollion (2018b), while McHugh's (2022; 2023) aboutness approach accounts for our data.

2 Lewis's and Stalnaker's Semantics Validate Reciprocity

(1) **Lewis's Semantics** (Lewis 1981: 230)

a. For each world w, let $\leq_w$ be a reflexive and transitive binary relation over the set of possible worlds.[3]

b. $A > C$ is true at w just in case for every A-world x, there is an A-world $y \leq_w x$ such that for every world $z \leq_w y$, if A is true at z, C is true at z.

This semantics validates Reciprocity (see the Appendix for a proof). One may impose further constraints on the similarity order, such as strong centering or the limit assumption.[4] Reflexivity and transitivity are the absolute minimum constraints to impose on the order; without them the approach would be deeply unworkable. Moreover, if we construct the order from an ordering source as Kratzer (1981a: 47) proposes, in (2), then reflexivity and transitivity of the order follow, respectively, from the reflexivity and transitivity of implication—bedrock principles of logic.

(2) a. An ordering source g is a function from worlds to sets of propositions.

b. $w' \leq_w w''$ just in case for all $p \in g(w)$, if p is true at w'', p is true at w'.

Given that the order is reflexive and transitive, the truth conditions in (1b) above validate Reciprocity. The validity of Reciprocity therefore serves as a fundamental test for the correctness of Lewis's semantics, regardless of how the order is interpreted.

[3] Lewis (1981: 220) himself begins with a irreflexive order $<_w$ and constructs a reflexive order $\leq_w$ by taking $w' \leq_w w''$ just in case $w' <_w w''$ or $w' = w''$. Nonetheless, he states the semantics in (1b) in terms of $\leq_w$, which is reflexive and transitive.

[4] Under the limit assumption, the semantics in (1b) is equivalent the more familiar requirement that for every closest world to w where A is true, C is true; for formal details regarding the definition of the closest worlds and the limit assumption, see Kaufmann 2017.

(3) **Stalnaker's Semantics** (Stalnaker 1968: 35–36)

a. Let f be a function taking a proposition (a set of worlds) and a world and returning a world, subject to the following constraints.

(i) A is true at $f(A,w)$.

(ii) $f(A,w)$ is the absurd world (where every proposition is true) only if there is no possible world with respect to w in which A is true.

(iii) If A is true in w then $f(A,w) = w$.

(iv) If A is true in $f(B,w)$ and B is true in $f(A,w)$, then $f(A,w) = f(B,w)$.

b. $A > C$ is true at world w just in case C is true at $f(A,w)$.

Condition (iv) corresponds to Reciprocity. It says that if $A > B$ and $B > A$ are true at a world w, then the selected A-world at w is the selected B-world at w. A fortiori, C is true at the selected A-world just in case it is true at the selected B-world.

Stalnaker appears to take these constraints to be a non-negotiable feature of his account, remarking that "the selection is based on an ordering of possible worlds with respect to their resemblance to the base world. If this is correct, then [(iii) and (iv)] must be imposed on the *s*-function [the selection function]. ... These conditions on the selection function are necessary in order that this account be recognizable as an explication of the conditional" (Stalnaker 1968: 36).

3 Previous Counterexamples to Reciprocity

There are a number of previous alleged counterexamples to Reciprocity, including by Mårtensson (1999), Tooley (2002), and Ahmed (2011). Here we discuss two, which we take to be representative of previous counterexamples to Reciprocity.

3.1 Stalnaker's Jewel Case

Stalnaker (1984: 130) offers the following apparent counterexample to Reciprocity, modeled on one by Tichỳ (1978).[5]

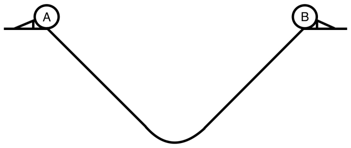
> The crown jewels are on an open display platform surrounded by electric eye sensors. A cat is sleeping on the platform, near the jewels but outside the circle of electric eyes. If anyone, human or cat, were to reach into the display area, an alarm would sound. If the alarm were to sound, it would wake up the cat. If the cat were to wake up, he would cross into the display area, setting off the alarm.

[5] We do not discuss Tichỳ's example here, since Stalnaker's has the same structure but is much clearer.

(4) a. If the alarm had sounded, the cat would have woken up.
b. If the cat had woken up, the alarm would have sounded.
c. If the cat had woken up, he would have set off the alarm.
d. If the alarm had sounded, the cat would have set it off.

Stalnaker replies that the three premises are true relative to different contexts. (4a) is only true when we ignore the possibility of the cat setting off the alarm, while for (4c) to be true we must take this possibility into account.[6] We agree with Stalnaker here, and add that on our diagnosis, this case illustrates the possibility of *backtracking*. We may ignore or take into account what would have lead to the alarm sounding.[7] If we ignore the possibility of the cat setting it off, (4a) is true but (4c) false. If we consider it, (4a) is false and (4c) true. On either interpretation, then, at least one premise is false, and the case does threaten the validity of Reciprocity.

3.2 Bacon's Hill



Bacon (2013) presents the following scenario, illustrated on the right. Two balls, A and B, are balanced on opposing slopes. If A falls, it rolls down the hill and up the other side, causing B to fall. Similarly, if B falls it rolls down and up again, causing A to fall. A and B each sit on a pressure plate. If the pressure is released on A's plate while B's is still depressed, a light will turn green and remain green. If the pressure is released on B's plate while A's is still depressed, the light will turn red and remain red. The light cannot be both green and red at once. Currently neither ball has fallen and the light is off. Consider the following instance of Reciprocity.

(5) a. If A fell, B would fall.
b. If B fell, A would fall.
c. If A fell, the light would turn green.
d. If B fell, the light would turn green.

[6] Stalnaker (1984: 132) puts it like this. "You say, 'If the alarm had sounded, it would have woken up the cat.' Not so, I reply, 'The cat would have been already awake. For if the alarm had sounded, it would have been the cat who set it off (no one else was around to set it off), and if the alarm had sounded because the cat set it off, it wouldn't have been what woke up the cat'."

[7] Recent experimental work has found significant variability in whether or not people backtrack. Rips (2010) found that people sometimes backtrack, Sloman & Lagnado (2005) found greater evidence for non-backtracking, interventionist interpretations that leave the past unchanged, and Gerstenberg, Bechlivanidis & Lagnado (2013) found that participants are more likely to backtrack when explicitly asked to consider the causes of the antecedent being true. It is therefore natural to expect that both forward-tracking and backtracking readings are available for (4).

Bacon acknowledges the existence of backtracking readings, whereby we consider the possibility that if A fell, it could be because B hit it, in which case the light would be red. He responds, “We only need [(5a–c)] to be simultaneously true in one context to complete our case against CSO [Reciprocity], which deems them jointly inconsistent.” We agree with Bacon here: there is forward-tracking interpretation on which the premises (5a–c) are all true and the conclusion (5d) is false. Indeed, it has often been observed that forward-tracking readings are the default, which would imply that the premises are true and the conclusion false on the *default* reading.[8]

There is a crucial difference between Stalnaker’s and Bacon’s cases. In Stalnaker’s, the premises are true relative different interpretations, while in Bacon’s they are true relative to the same interpretation; namely, the forward-tracking reading. Stalnaker’s scenario therefore does not threaten Reciprocity, but Bacon’s does. It is impossible to construct a model of Lewis’s or Stalnaker’s framework in which the premises (5a–c) are true but the conclusion (5d) is not. Given that such a reading exists, Bacon’s scenario appears to suggest that Reciprocity is intuitively invalid.

There is, however, a reply to Bacon’s scenario which may disqualify it as a counterexample to Reciprocity.[9] So far we have abstracted away the temporal information in the antecedent and consequent. One might propose that conditional antecedents and consequents are evaluated at different times; for example, Condoravdi (2002) argues that modals—such as *would*—expand the local time of evaluation into the future. This suggests that a sentence such as *A fell* may have different evaluation times when it appears in a conditional antecedent or consequent, represented in (6).

(6) a. If A fell [at time t], B would fall [at some time $t' \geq t$].
b. If B fell [at time t], A would fall [at some time $t' \geq t$].
c. If A fell [at time t], the light would turn green [at some time $t' \geq t$].
d. If B fell [at time t], the light would turn green [at some time $t' \geq t$].

If we take the evaluation time as part of the identity-criteria for sentences, then *A fell at t* and *A fell at some time* $t' \geq t$ are properly regarded as distinct sentences (similarly for B). On this assumption, Bacon’s counterexample in (5) in fact has the form, $A > B'$, $B > A'$, $A > C$, $B > C$. This is not an instance of Reciprocity, and can be captured by the similarity approach on Lewis’s or Stalnaker’s semantics.

[8] Lewis (1979: 457) writes that, “A counterfactual saying that the past would be different if the present were somehow different may come out true under the special resolution of its vagueness, but false under the standard resolution.” Khoo (2017) proposes, and aims to account for the fact that “The default interpretation of a forward counterfactual is non-backtracking.”

[9] We are grateful to Paul Egré (personal communication) and members of the Linguae Seminar at Institut Jean-Nicod for raising this possibility.

3.3 Similarity in Time

Let us suppose that the above objection can be met: Bacon's scenario is indeed a counterexample to Reciprocity. One might be tempted to conclude that the concept of an order, intuitively representing similarity, has no role to play in the semantics of conditionals. There is, however, a family of approaches to conditionals that provide a more limited role for similarity and which can model Bacon's scenario. The idea is to make similarity relative to a time.[10] This is in contrast to Lewis (1973), who took the order to represent a time-independent notion of overall similarity.

A representative of this view is Bennett (1984), who proposes that a counterfactual is true just in case the consequent holds at all worlds which, among those where the antecedent is true and that obey the laws of the actual world, are closest to the actual world *at the time to which the antecedent pertains*. This proposal has the resources to account for Bacon's example. It is natural to suppose the closest world at time t where A falls is one where B is still stationary at t. In that case the light turns green and sometime later B falls. Similarly, the closest world at time t where B falls is one where A is still stationary at t, the light turns red and sometime later A falls. Thus on Bennett's theory (5a–c) come out true while (5d) comes out false.

3.4 Interim Summary

Taking stock, Stalnaker's jewel case is not a counterexample to Reciprocity, since the premises are not jointly true with respect to the same reading. While Bacon's scenario overcomes this problem, it also does not provide convincing evidence against the similarity approach to conditionals. This is for two reasons. Firstly, one may argue that once we attend to temporal information in antecedents and consequents, it is no longer an instance of Reciprocity. Secondly, one may make the notion of similarity time-relative, thereby accounting for the case while nonetheless preserving a role for similarity in the semantics of conditionals.

We may call these alleged counterexamples to Reciprocity *dynamic counterexamples*, since time plays an essential role in their interpretation. In what follows we present novel, *static counterexamples* to Reciprocity. By circumventing the role of time, our counterexamples do not face the same objections as Bacon's scenario, and cannot be modelled by making similarity relative to a time. Even this more limited role for similarity will turn out to be misguided.

[10] Work in conditionals incorporating time-relativity includes Thomason 1970, Slote 1978, Lewis 1979, Thomason & Gupta 1980, Arregui 2005, Ippolito 2013, Khoo 2015, and Canavotto 2020.

4 The Experiment

We conducted an experiment to test whether Reciprocity is intuitively valid.[11] Thus we tested whether it is always the case that whenever a participant accepts the premises $A > B$, $B > A$ and $A > C$, they accept the conclusion, $B > C$.

4.1 Methods

Our design follows Ciardelli, Zhang & Champollion (2018b), Schulz (2018), and Romoli, Santorio & Wittenberg (2022), using a picture-sentence verification task. We improved on the design by testing three different scenarios, thus increasing the number of target trials per participant and allowing us to compare whether participants used the same interpretative strategy across multiple trials.

4.1.1 Participants

We tested the judgements of 80 native English speakers recruited via Prolific (taking only those with an approval rate above 90%). We excluded two participants from the analysis for not meeting the predetermined threshold of at most 25% error rate.

Since we created novel materials, we conducted a pilot study to asses the difficulty of the materials and the sample size. For the pilot we recruited 20 participants using the same criteria. Pilot participants were prohibited from taking the main study. The participants were informed about their rights and that the study was approved by the Ethics Committee of the Faculty of Humanities of the University of Amsterdam (FGW-1461). Participants were paid £1.75 for their participation and the median time it took them to complete the experiment was 9 minutes and 26 seconds.

4.1.2 Materials

An instance of Reciprocity consists of three premises and a conclusion. We designed three scenarios to test its validity. Here we present one of the scenarios we tested. The participants saw Figure 1, accompanied by the following instructions.

> There are two switches, A and B, connected to a light. Part of the circuit is shaded grey. Each switch has three possible positions: *up*, *in the middle*, or *down*. The current position of the switch is indicated by a green circle.
>
> As the wiring indicates, the light is on just in case switch A is in the middle and switch B is either up or in the middle.

[11] You can try the experiment for yourself at https://www.tklochowicz.com/experiment_reciprocity.

Currently, switch *A* is in the middle and switch *B* is down, so the light is off.

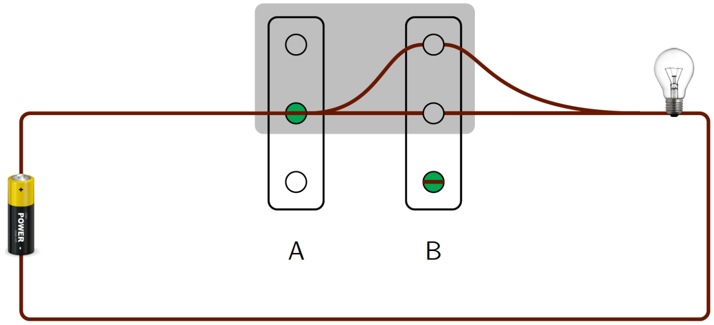


Figure 1 The switches scenario, one of three scenarios we tested.

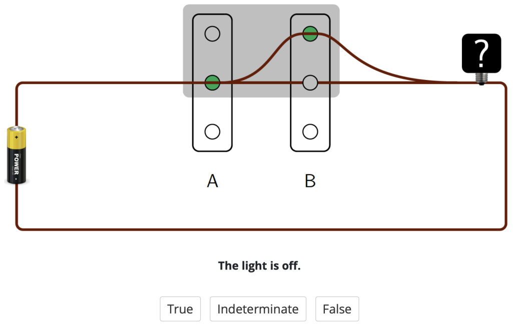
Participants saw the premises (P1–P3), the conclusion (C), a True (T) and a False (C) control, listed in (7), as well as six fillers, presented in random order.

(7) T If both switches were in the middle, the light would be on.
P1 If switch B were in the shaded area, both switches would be in the shaded area. $B > \mathit{both}$
P2 If both switches were in the shaded area, switch B would be in the shaded area. $\mathit{both} > B$
P3 If switch B were in the shaded area, the light would be on. $B > \mathit{on}$
C If both switches were in the shaded area, the light would be on. $\mathit{both} > \mathit{on}$
F If both switches were outside the shaded area, the light would be on.

Following Romoli et al. (2022), for each sentence we asked participants whether it is true, false, or indeterminate. If they answered indeterminate we asked them to choose between "I strongly feel that there is no right answer" and "I just don't know; I am not sure what the answer is". We added an indeterminate option since some theories (e.g. Stalnaker's 1980 supervaluationism) allow conditionals to be neither true nor false.

Our two other scenarios had the same logical structure, but described different situations. In the *candy* scenario the participants were told what flavour of candy two children got, and had to decide whether both children would be happy with their candy if the distribution was different. In the third scenario (*numbers*) the task was similar but concerned cards dealt to players of a game.[12]

[12] All three scenarios can be accessed via the online version of the experiment; see footnote 11.



The light is off.

Figure 2 One of four questions from the training phase of the switches scenario.

4.1.3 Procedure

We designed the experiment using the jsPsych library by de Leeuw, Gilbert & Luchterhandt (2023), and published it through *cognition.run*. At the outset, the participants saw information about the experiment and were asked to express their informed consent for participation. They were then asked to confirm that they were a native speaker of English.

The experiment consisted of three blocks, one for each scenario (*switches*, *candy* and *numbers*), presented in a random order. Each block began with a training phase to familiarise the participants with the materials and the rules in each scenario. They had to answer four questions which tested their understanding of the rules. Figure 2 illustrates one of the four questions asked during the training phase for the switches scenario. The participants received feedback on each answer during the training phase. If the answer was correct the word "*Correct!*" appeared briefly on the screen. If the answer was incorrect, the word "*Incorrect!*" appeared alongside the information that they had to wait four seconds and try again. Participants could not proceed unless they gave correct answers to all training trials. The participants were then asked to judge the test sentences, controls (see (7)), and six fillers.[13] All sentences were presented in a random order.

4.1.4 Data treatment and statistical analysis

The data were collected using *cognition.run*, downloaded as a JSON file and converted into a *.csv* using python. To perform the statistical analysis, we used language *R*, the *lme*4 library and other libraries (R Core Team 2021; Bates, Mächler, Bolker

[13] There were four false fillers and two true fillers. This was to result in an even distribution of true and false answers, assuming that the premises are true in the given contexts.

& Walker 2015). We visualised our data using *ggplot2* by Wickham (2016).[14]

We excluded from the statistical analysis all responses in which the participant selected indeterminate and then chose the response “I just don’t know; I am not sure what the answer is.” Since we were interested in whether Reciprocity is truth-preserving, we treated the other indeterminate response (“I strongly feel that there is no right answer”) as not true. We used mixed effect logistic regression as a model for acceptance rates (*true* versus *false or indeterminate*). We considered two separate GLMER models: the first had the true control (T) as a baseline and the second the conclusion (C).

We used linear regression to analyse reaction times. The reaction times were centered around the each participant’s mean reaction time to accurately represent whether the participant spent more (positive value) or less (negative value) time in a given trial than on an average trial.

4.2 Results

4.2.1 The pilot

We piloted the study to evaluate the difficulty of the scenarios. We concluded that the participants could understand the task well since the error rates for the filler items, as well as the controls did not exceed 25%. The acceptance rates for the premises (P1–P3), which was above 85%, allowed us to estimate the sample size needed to get meaningful results (i.e. to have enough trials where all premises are accepted).

4.2.2 Acceptance rates

Participants understood our scenarios well, with a mean accuracy of 90% on the filler items, confirming the results of the pilot. Moreover, high accuracy on the control items (see below) indicate that the participants were able to understand and correctly process sentences of similar complexity to the premises and the conclusion.

Table 1 and Figure 3 present the participants’ responses from all three scenarios. In the first model (GLMER) we found no significant difference in rates of true judgements between the *True* control and any of the premises *P1–P3* (all $\beta \approx 2\%, \text{all } p > 0.15$). However, the conclusion C was significantly different from the true control ($\beta \approx 24\%, \text{all } p < 0.001$). In the second statistical model (GLMER), we found out that all the conditions were significantly different from the conclusion (all $|\beta| > 22\%, p < 0.001$). Analysing each scenario individually revealed effects of similar size and significance. We did not find any significant interaction between conditions (T vs C) and scenarios (all $\beta \approx 3\%$, all $p > 0.02$).

[14] The data and the statistical analysis can be found at https://osf.io/v765x/.

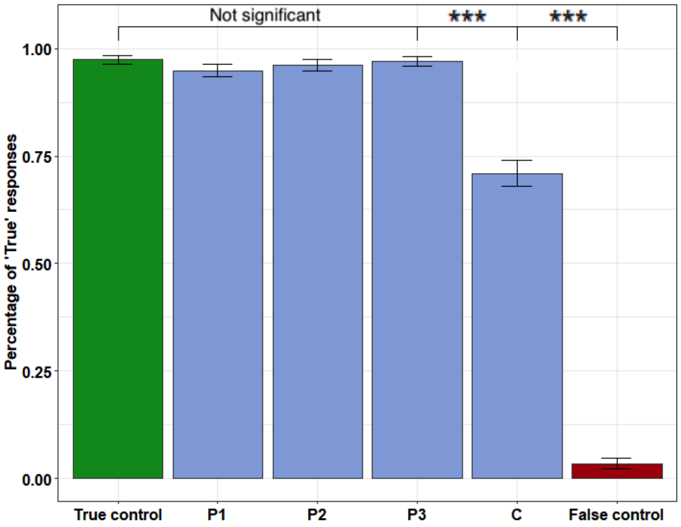


Figure 3 Percentage of ‘True’ responses. Error bars denote Standard Errors.

Sentence	True	Indeterminate	False	Not sure
True control	232	0	8	0
Premise 1	225	0	11	4
Premise 2	227	2	8	3
Premise 3	225	2	12	1
Conclusion	163	26	43	8
False control	9	0	229	2

Table 1 Responses from all three scenarios.

On average, in 28% of cases where a participant accepted all of the premises P1–P3 they rejected the conclusion C. Moreover, half (39/78) of the participants exhibited this behaviour at least once in a scenario where they additionally also answered all the controls correctly. Figure 4 presents how participants’ strategies changed during the experiment.

4.2.3 Reaction Times

Participants responded to the trials with a mean reaction time of 6.8 seconds and a standard deviation of 4.7 seconds. We excluded 25 outlier trials, which were longer than 21 seconds (3 standard deviations from the mean). Participants took similar amount of time to react to each scenario ($|\beta| < 0.3$sec; $p \approx 0.04; 0.23$ for *candy*

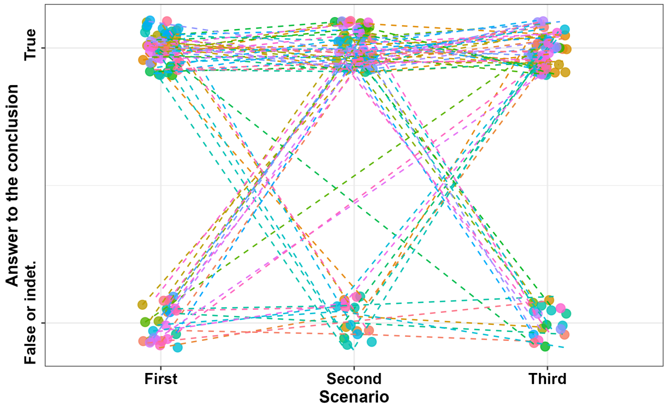


Figure 4 Change in participants' strategies in evaluation of the conclusion for different scenarios, across the experiment. Diagonal lines correspond to a change in response and horizontal lines indicate lack of change.

and *switches* and $\beta \approx 0.6$sec; $p \approx 0.001$ for *numbers*). On average participants reacted the slowest to Premise 2 ($\beta \approx 1.7$sec), but they were fast with the controls ($\beta \approx -0.7$sec for False and $\beta \approx -1.2$sec for True; all $p < 0.001$). The other premises and the conclusion were not different from the average reaction time ($\beta \approx 0.2$sec; $p \approx 0.2$ for P1 $\beta \approx 0.3$sec; $p \approx 0.07$ for P3, and $\beta \approx 0.4$sec; $p \approx 0.06$ for C). The reaction times are represented in Figure 5.

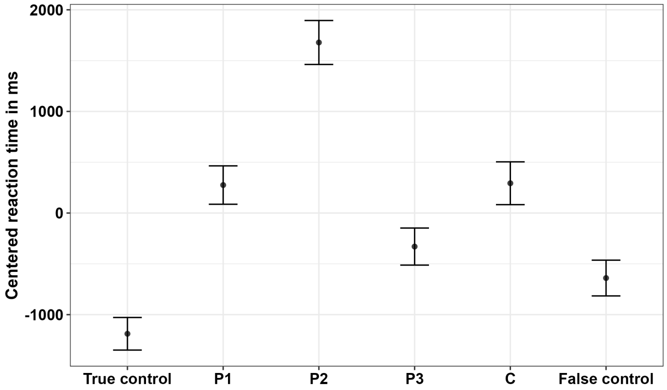


Figure 5 Centered reaction times to all test trials. Error bars denote Standard Errors.

In the second linear regression, we tested the differences between responses to the Conclusion. When a participant gave the answer false to such a trial, they spent on average 1.9sec longer than if they were to answer true ($p < 0.001$). They also spent longer (1.7sec) if they answered indeterminate, but this difference is barely significant ($p \approx 0.005$), since the number of those responses was small. Moreover we found a significant interaction between answer polarity (*true* vs. *false or indeterminate*) and the type of condition (*control* vs. *conclusion*; $p < 0.001$).

4.3 Discussion

Our results indicate that Reciprocity is invalid. The premises can all be judged to be true while the conclusion is not. In the scenarios we tested, the premises were all accepted at ceiling, with the conclusion accepted significantly less frequently. Half of the participants judged in at least one scenario that the premises are true but conclusion is not (42/78). This remains (39/78) when we restrict to those who correctly answered the true and false controls, indicating that they understood the scenario and reasoned correctly with conditionals of equal complexity to the test items. Overall, the inference rule was violated—the premises were all judged true but the conclusion was not—in 28% of test trials.

Our results nonetheless indicate that the conclusion is not categorically false. It was judged to be true 70% of the time. Hence, to explain our data, any theory of conditionals must allow for some variability in the status of the Conclusion.

Turning to reaction times, we observed that evaluating the conclusion as false or indeterminate took longer than evaluating it as true. This may be because, in the latter case, the participants only considered moving switch B. Since switch A satisfies the condition of being in the shaded area, imagining it in a different position may be more costly.

Alternatively, one might be tempted to propose that as a general rule, it takes longer to reject a statement than to accept it (for discussion see e.g. Dudschig, Kaup, Liu & Schwab 2021), and that this explains why evaluating the conclusion as false or indeterminate took longer than evaluating it as true. However, note that the difference between accepting and rejecting the conclusion is significantly greater than the contrast between the true and false controls. This suggests that a general delay for rejection cannot alone account for why it took longer to reject the Conclusion rather than accept it.

Moreover, we discovered that verifying the analytically true statement P2 took much longer. This may be a result of pragmatic oddness associated with verifying non-informative statements (Grice 1975).

5 General Discussion

Our results offer a novel and striking challenge to the similarity approach to conditionals. It is impossible to construct a model of Lewis's or Stanlaker's semantics in which premises are true but the conclusion is not. However, a significant number of participants reasoned in this way.

On the basis of our results, we propose abandoning the similarity approach. The question is what alternative framework to adopt in its place. Here we discuss four proposals: Kratzer (2012), Fine (2012), Ciardelli et al. (2018b), and McHugh (2023).

5.1 Kratzer's Lumping Semantics of Counterfactuals

In this section we show that Kratzer's semantics of counterfactuals (Kratzer 1989, 2012), on the most natural modelling of the switches scenario, predicts the Conclusion to be true. This semantics therefore does not for the significant number of participants who judged that the Conclusion is not true.

Kratzer (1989) works in a situation semantics, where situations are parts of possible worlds and propositions defined as sets of situations (as usual, we take the elements of a proposition to be those situations where the proposition is true). Kratzer defines the following relationship of *lumping* between propositions.

(8) **Kratzer's definition of lumping** (Kratzer 1989: 616)
Where p and q are sets of situations, and w a world, p lumps q at w iff

a. p is true at w, and
b. Every situation that is part of w where p is true, q is true.

Kratzer starts with what she calls a *base set* F_w, which is a set of propositions all of which are true in the world of evaluation w.[15] Kratzer (1989: 635; 2012: 133) proposes the following semantics of "would"-conditionals.

(9) For any world w, admissible base set F_w, and proposition p, the *crucial set* $F_{w,p}$ is the set of all subsets A of $F_w \cup \{p\}$ satisfying the following conditions.
(i) A is consistent.
(ii) $p \in A$.
(iii) A is closed under lumping: for all $q \in A$, $r \in F_w$, if q lumps r in w, $r \in A$.

(10) Given a world w and an admissible Base Set F_w, a "would"-counterfactual with antecedent p and consequent q is true in w iff for every set in $F_{w,p}$ there is a superset in $F_{w,p}$ that logically implies q.

[15] Kratzer (2012: 132–33) imposes some constraints on what counts as an admissible base set which we will not discuss here. As the reader may verify, they are all satisfied in the choice of base set to come.

Let us apply this framework to the Conclusion of our switches scenario, “If both switches were in the shaded region, the light would be on.” The base set represents the basic facts about actual scenario. The most natural choice of base set in this scenario is $F_w = \{$*Switch A is in the middle*, *Switch B is down*, *The light is on iff A is in the middle and B is up or in the middle*, *A switch is in the shaded region iff it is up or in the middle*$\}$.[16] Given this base set, Kratzer predicts the Conclusion to be true.[17]

Stepping back, it is natural to expect that Kratzer’s approach predicts the conclusion to be true. Throughout her extensive work on the semantics of counterfactuals, she has told a detailed and remarkably compelling story about their meaning:

> The truth of counterfactuals depends on everything which is the case in the world under consideration: in assessing them, we have to consider all the possibilities of adding as many facts to the antecedent as consistency permits. If the consequent follows from every such possibility, then (and only then), the whole counterfactual is true.
>
> (Kratzer 1981b: 201)[18]

The antecedent at issue is “Both switches are in the shaded area.” The fact that switch A is in the middle is always consistent with both switches being in the shaded area (since the middle position is inside the shaded area). Thus the fact that switch A is in the middle can always be added to the antecedent while preserving consistency. It is therefore not only the letter, but also the spirit of Kratzer’s approach which predicts the Conclusion to be true, and which does not account for the significant

[16] The reader may wonder why *The light is off* is not in the base set. The reason is that including it would generate clearly incorrect predictions for other conditionals, including our true control, “If both switches were in the middle, the light would be on.” To see this, suppose that *The light is off* is in the base set. Consider $A = \{$*Both switches are in the middle*, *A is in the middle*, *The light is off*$\}$. This satisfies (i–iii), and is therefore in the crucial set. However, no more elements of the base set can be added to A while maintaining consistency: adding *B is down* or *The light is on iff A is in the middle and B is the middle or up* results in inconsistency. So every superset of A satisfying (i–iii) (namely, only A itself) logically implies that the light is off. Kratzer’s semantics would therefore predict the true control to be false. Removing *The light is off* from the base set solves the problem.

[17] For concreteness, the crucial set is given below. Note that the final element is a superset of every element and logically implies that the light is on. Kratzer therefore predicts the Conclusion to be true.

$\{\{$*Both in shaded*$\}$,
$\{$*Both in shaded*, *A mid*$\}$,
$\{$*Both in shaded*, *Light on iff A mid and B up or mid*$\}$,
$\{$*Both in shaded*, *Switch in shaded iff up or mid*$\}$,
$\{$*Both in shaded*, *A mid*, *Light on iff A mid and B up or mid*$\}$,
$\{$*Both in shaded*, *A mid*, *Switch in shaded iff up or mid*$\}$,
$\{$*Both in shaded*, *Light on iff A mid and B up or mid*, *Switch in shaded iff up or mid*$\}$,
$\{$*Both in shaded*, *A mid*, *Light on iff A mid and B up or mid*, *Switch in shaded iff up or mid*$\}\}$

[18] A similar statement also appears in Kratzer (1989: 626; 2012: 127).

number of participants who judged that it is not true.

5.2 Fine's Truthmaker Semantics of Counterfactuals

Fine (2012) proposes a semantics of *would*-conditionals using truthmaker semantics. This framework is centered around the notion of *exact verification*. The guiding idea is that a part of the world, or *state*, exactly verifies a sentence just in case the state's obtaining is wholly relevant to the truth of the sentence (Fine 2017). Fine (2012) offers the following semantics of *would*-conditionals in terms of exact verification.

> $A > C$ is true at a world w just in case for every exact verifier t of A and possible outcome u of t at w, u contains an exact verifier of C.

Note that this semantics satisfies a principle Fine calls *Universal Realisability of the Antecedent*: a conditional is true only if it is true for any way in which its antecedent might be true (Fine 2012: 236). Plausibly, one way for both switches to be in the shaded area is for them to both be up, in which case the light is off. Fine's semantics therefore leads us to expect the Conclusion, "If both switches were in the shaded area the light would be on" to be false.

More formally, let us represent the antecedent as a conjunction, *Switch A is in the shaded area and switch B is in the shaded area*. Fine (2017: 562) offers an exact verification clause for conjunction: a state s exactly verifies $A \wedge B$ just in case s is the fusion of states t and t', where t exactly verifies A and t' exactly verifies B. It is natural to assume that the state of switch A being up exactly verifies that switch A is in the shaded area; similarly for switch B. Thus their fusion exactly verifies that both switches are in the shaded area. One outcome of both switches being up is that the light is off. Fine therefore predicts the Conclusion to be false. This does not account for the majority of participants who judged it true.

5.3 Ciardelli, Zhang & Champollion's background semantics

Ciardelli et al. (2018b) work within inquisitive semantics, where a proposition is a set of information states, and an information state a set of worlds (Ciardelli, Groenendijk & Roelofsen 2018a). Plausibly, we assume that the antecedent "Both switches are in the shaded area" is not inquisitive (i.e. consists of a single information state), so this aspect of their proposal does not affect their predictions for our experiment.

They offer a semantics of counterfactuals based on the intuitive idea that counterfactual antecedents draw a distinction between foreground and background facts. "Background facts are held fixed while making a counterfactual assumption, while foreground facts are allowed to change" (2018b: 599). They propose that a fact is allowed to change when it is "called into question" by the conditional antecedent.

This happens when the fact either (i) contributes to the falsity of the antecedent in the actual world, or (ii) is causally dependent on a such a fact. A fact f contributes to the falsity of a proposition a at world w just in case there is some set of facts F at w that is consistent with a, but $F \cup \{f\}$ is not.

Ciardelli et al. frame their account within structural causal models.[19] We assume that the causal model of our switches scenario consists of a variable representing whether switch A is up, in the middle, or down, a similar variable for switch B, one representing whether the light is on or off, and an equation stating that the light is on just in case switch A is in the middle and B is in the middle or up.[20]

Given these assumptions, Ciardelli et al. predict that we fix the fact that switch A is in the middle when we interpret the antecedent “If both switches were in the shaded area”. This fact does not contribute to the falsity of this antecedent (since switch A being in the middle is always consistent with both switches being in the shaded area), nor is it causally dependent on such a fact. Thus their proposal, like the similarity approach and Kratzer’s semantics, predicts the conclusion to be true.

5.4 McHugh’s Aboutness Semantics of Counterfactuals

McHugh proposes the following semantics of conditionals, based on a notion of aboutness (for a formalisation see McHugh 2022, 2023: chapter 2).

1. Pick a time at which to imagine the antecedent true.
2. Allow the part of the world the antecedent is about at that time to vary.
3. Play forward the laws.
4. Restrict to worlds where the antecedent is true.
5. Check if the consequent is true at the resulting world(s).

We write “world(s)” to cover selectional approaches to conditionals, according to which a conditional is true just in case the consequent is true at the unique world selected from the relevant domain (e.g. Cariani & Santorio 2018; Mandelkern 2018).

The thought is that when we interpret a conditional, we allow the part of the world its antecedent is *about* to vary. The guiding idea of aboutness is that a sentence is about the part of the world that exactly determines its truth value (McHugh 2023: 108). This allows for variation in what a sentence is about; for example, McHugh (2023: 123–24) offers two views on aboutness. On the truthmaker view, a sentence is about the parts of the actual world that exactly verify or falsify it. The state of switch B, but not switch A, exactly verifies or falsifies that both switches are in the

[19] See for example Pearl 2000, Schulz 2007, and Kaufmann 2013.

[20] Though note that Ciardelli et al.’s (2018b) predictions would remain the same if we replaced these variables with ones merely representing whether or not each switch is in the shaded area.

shaded area, so this view predicts that participants only vary the position of switch B when interpreting the conclusion. The truthmaker view therefore predicts the conclusion to be true.

On the subject matter view, a sentence is about the exact verifiers and falsifiers of its *atomic* sentences.[21] We assume, plausibly, that *switch A is in the shaded area* is an atomic sentence of *both switches are in the shaded area*. This predicts that we vary switch A when we interpret C. The subject matter view therefore predicts the conclusion to be false. Assuming that our concept of aboutness is vague, allowing for both of these views, correctly predicts the intermediate status of the conclusion.

Appendix. Proof that Lewis's semantics in (1) validates Reciprocity

Proof. (See Figure 6.) For any world w and sentence A, let $w \models A$ denote that A is true at w. Pick any world w and suppose $A > B$, $B > A$ and $B > C$ are true at w. To show that $A > C$ is true at w, pick any $x \models A$. We have to show that there is a $y \models A$ such that $y \leq_w x$ and for all $z \leq_w y$, $z \models A \rightarrow C$, where $\rightarrow$ is the material conditional.

Since $w \models A > B$ and $x \models A$, there is a $v \models A$ such that $v \leq_w x$ and (i) for all $v' \leq_w v$, $v' \models A \rightarrow B$. Since $\leq_w$ is reflexive, $v \leq_w v$, so $v \models A \rightarrow B$. Thus $v \models B$.

Since $w \models B > A$ and $v \models B$, there is a $u \models B$ such that $u \leq_w v$ and (ii) for all $u' \leq_w u$, $u' \models B \rightarrow A$. Since $w \models B > C$ and $u \models B$, there is a $y \models B$ such that $y \leq_w u$ and (iii) for all $z \leq_w y$, $z \models B \rightarrow C$. Since $y \leq_w u$, by (ii), $y \models B \rightarrow A$. Then as $y \models B$, $y \models A$. And as $y \leq_w u \leq_w v \leq_w x$, by transitivity of $\leq_w$, $y \leq_w x$.

We show that $z \models A \rightarrow C$ for all $z \leq_w y$. Pick any $z \leq_w y$. Then $z \leq_w y \leq_w u \leq_w v$, so by transitivity of $\leq_w$, $z \leq_w v$. Then by (i), $z \models A \rightarrow B$. And since $z \leq_w y$, by (iii), $z \models B \rightarrow C$. Hence $z \models A \rightarrow C$. □

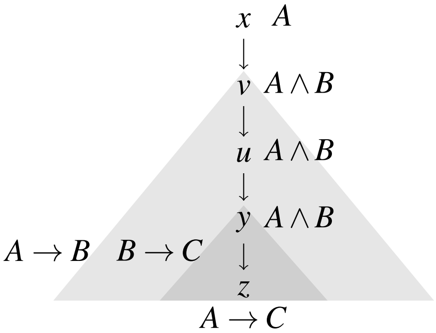


Figure 6 Illustrating the proof that Reciprocity is valid on Lewis's semantics.

[21] This follows what Hawke (2018) calls an 'atom-based' approach to subject matter. For discussion of the notion of subject matter see e.g. Yablo 2014: 42, Fine 2016: 11, and Berto 2018: 1878.

References

Ahmed, Arif. 2011. Walters on conjunction conditionalization. In *The Aristotelian Society*, vol. 111, 115–122. doi:10.1111/j.1467-9264.2011.00301.x.

Arregui, Ana. 2005. *On the accessibility of possible worlds: The role of tense and aspect*: University of Massachusetts Amherst PhD dissertation. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.456.7358&rep=rep1&type=pdf.

Bacon, Andrew. 2013. In defence of a naïve conditional epistemology. *Manuscript* .

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. doi:10.18637/jss.v067.i01.

Bennett, Jonathan. 1984. Counterfactuals and temporal direction. *The Philosophical Review* 93(1). 57–91. doi:10.2307/2184413.

Berto, Francesco. 2018. Aboutness in imagination. *Philosophical Studies* 175(8). 1871–1886. doi:10.1007/s11098-017-0937-y.

Canavotto, Ilaria. 2020. *Where responsibility takes you: Logics of agency, counterfactuals and norms*: University of Amsterdam PhD dissertation. doi:10.1007/978-3-031-17111-6.

Cariani, Fabrizio & Paolo Santorio. 2018. *Will* done better: Selection semantics, future credence, and indeterminacy. *Mind* 127(505). 129–165. doi:10.1093/mind/fzw004.

Ciardelli, Ivano, Jeroen Groenendijk & Floris Roelofsen. 2018a. *Inquisitive Semantics* Oxford Surveys in Semantics and Pragmatics. Oxford University Press. doi:10.1093/oso/9780198814788.001.0001.

Ciardelli, Ivano, Linmin Zhang & Lucas Champollion. 2018b. Two switches in the theory of counterfactuals. *Linguistics and Philosophy* 41(6). 577–621. doi:10.1007/s10988-018-9232-4.

Condoravdi, Cleo. 2002. Temporal interpretation of modals: Modals for the present and for the past. In *The Construction of Meaning*, 59–88. CSLI Publications. https://semanticsarchive.net/Archive/2JmZTIwO/temp-modals.ps.

Dudschig, Carolin, Barbara Kaup, Mingya Liu & Juliane Schwab. 2021. The processing of negation and polarity: An overview. *Journal of Psycholinguistic Research* 50(6). 1199–1213.

Egré, Paul & Hans Rott. 2021. The Logic of Conditionals. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University Winter 2021 edn.

Fine, Kit. 2012. Counterfactuals without possible worlds. *Journal of Philosophy* 109(3). 221–246. doi:10.5840/jphil201210938.

Fine, Kit. 2016. Angellic content. *Journal of Philosophical Logic* 45(2). 199–226. doi:10.1007/s10992-015-9371-9.

Fine, Kit. 2017. Truthmaker semantics. In *A Companion to the Philosophy of Language*, chap. 22, 556–577. Wiley Blackwell. doi:10.1002/9781118972090.ch22.

Gerstenberg, Tobias, Christos Bechlivanidis & David A Lagnado. 2013. Back on track: Backtracking in counterfactual reasoning. In *Annual meeting of the Cognitive Science Society*, vol. 35 35, https://escholarship.org/uc/item/95w6b2bn.

Grice, Herbert Paul. 1975. Logic and conversation. In *Speech acts*, 41–58. Brill.

Hawke, Peter. 2018. Theories of aboutness. *Australasian Journal of Philosophy* 96(4). 697–723. doi:10.1080/00048402.2017.1388826.

Ippolito, Michela. 2013. *Subjunctive Conditionals: A Linguistic Analysis*. MIT Press.

Kaufmann, Stefan. 2013. Causal premise semantics. *Cognitive Science* 37(6). 1136–1170. doi:10.1111/cogs.12063.

Kaufmann, Stefan. 2017. The limit assumption. *Semantics and Pragmatics* 10. 18–1. doi:10.3765/sp.10.18.

Khoo, Justin. 2015. On indicative and subjunctive conditionals. *Philosophers' Imprint* 15(32). 1–40. http://hdl.handle.net/2027/spo.3521354.0015.032.

Khoo, Justin. 2017. Backtracking counterfactuals revisited. *Mind* 126(503). 841–910. doi:10.1093/mind/fzw005.

Kratzer, Angelika. 1981a. The notional category of modality. *Words, Worlds, and Contexts: New Approaches in Word Semantics. Berlin: W. de Gruyter* 39–74. doi:10.1515/9783110842524-004.

Kratzer, Angelika. 1981b. Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic* 10(2). 201–216. doi:10.1007/BF00248849.

Kratzer, Angelika. 1989. An investigation of the lumps of thought. *Linguistics and philosophy* 607–653. doi:10.1007/BF00627775.

Kratzer, Angelika. 2012. *Modals and conditionals*. Oxford University Press. doi:10.1093/acprof:oso/9780199234684.001.0001.

de Leeuw, Joshua R, Rebecca A Gilbert & Björn Luchterhandt. 2023. jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software* 8(85). 5351.

Lewis, David. 1973. *Counterfactuals*. Wiley-Blackwell.

Lewis, David. 1979. Counterfactual dependence and time's arrow. *Noûs* 455–476. doi:10.2307/2215339.

Lewis, David. 1981. Ordering semantics and premise semantics for counterfactuals. *Journal of philosophical logic* 217–234. doi:10.1007/BF00248850.

Mandelkern, Matthew. 2018. Talking about worlds. *Philosophical Perspectives* 32(1). 298–325. doi:10.1111/phpe.12112.

Mårtensson, J. 1999. Subjunctive conditionals and time: A defense of the classical approach.

McHugh, Dean. 2022. Aboutness and modality. *Proceedings of the 23rd Amsterdam*

Colloquium 194–206. doi:10.21942/uva.21739718.

McHugh, Dean. 2023. *Causation and Modality: Models and Meanings*: University of Amsterdam PhD dissertation. https://eprints.illc.uva.nl/id/eprint/2243.

Nute, Donald. 1980. *Topics in conditional logic*. Springer.

Pearl, Judea. 2000. *Causality*. Cambridge University Press.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

Rips, Lance J. 2010. Two causal theories of counterfactual conditionals. *Cognitive science* 34(2). 175–221. doi:10.1111/j.1551-6709.2009.01080.x.

Romoli, Jacopo, Paolo Santorio & Eva Wittenberg. 2022. Alternatives in Counterfactuals: What Is Right and What Is Not. *Journal of Semantics* 39(2). 213–260. doi:10.1093/jos/ffab023.

Schulz, Katrin. 2007. *Minimal models in semantics and pragmatics: Free choice, exhaustivity, and conditionals*: Institute for Logic, Language and Computation, University of Amsterdam PhD dissertation. https://www.illc.uva.nl/Research/Publications/Dissertations/DS-2007-04.text.pdf.

Schulz, Katrin. 2018. The similarity approach strikes back: Negation in counterfactuals. In Uli Sauerland & Stephanie Solt (eds.), *Sinn und Bedeutung 22*, vol. 2, 343–360. ZAS Papers in Linguistics. doi:10.21248/zaspil.61.2018.500.

Sloman, Steven A. & David A. Lagnado. 2005. Do we “do”? *Cognitive Science* 29(1). 5–39. doi:10.1207/s15516709cog2901_2.

Slote, Michael A. 1978. Time in counterfactuals. *The Philosophical Review* 87(1). 3–27. doi:10.2307/2184345.

Stalnaker, Robert. 1968. A theory of conditionals. In William L. Harper, Robert Stalnaker & Glenn Pearce (eds.), *Ifs*, 41–55. Springer. doi:10.1007/978-94-009-9117-0_2.

Stalnaker, Robert. 1984. *Inquiry*. MIT Press.

Stalnaker, Robert C. 1980. A defense of conditional excluded middle. In *Ifs: Conditionals, Belief, Decision, Chance and Time*, 87–104. Springer. doi:10.1007/978-94-009-9117-0_4.

Starr, W. 2022. Counterfactuals. In Edward N. Zalta & Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University Winter 2022 edn.

Thomason, Richmond. 1970. Indeterminist time and truth-value gaps. *Theoria* 36(3). 264–281. doi:10.1111/j.1755-2567.1970.tb00427.x.

Thomason, Richmond & Anil Gupta. 1980. A theory of conditionals in the context of branching time. In *Ifs*, 299–322. Springer. doi:10.1007/978-94-009-9117-0_15.

Tichỳ, Pavel. 1978. A new theory of subjunctive conditionals. *Synthese* 37(3). 433–457. doi:10.1007/BF00873249.

Tooley, Michael. 2002. Backward causation and the stalnaker-lewis approach to counterfactuals. *Analysis* 62(3). 191–197. doi:10.1093/analys/62.3.191.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Yablo, Stephen. 2014. *Aboutness*. Princeton University Press. doi:10.1515/9781400845989.

Dean McHugh
Science Park 107
1098 XG Amsterdam
deanmchugh1@gmail.com

Tomasz Klochowicz
Science Park 107
1098 XG Amsterdam
t.j.klochowicz@uva.nl