Two switches in the theory of counterfactuals

A study of truth conditionality and minimal change

Ivano Ciardelli
i.a.ciardelli@uva.nl

Linmin Zhang
linmin.zhang@nyu.edu

Lucas Champollion
champollion@nyu.edu

This draft compiled on November 6, 2016

Abstract

Based on a crowdsourced truth-value judgment experiment, we provide empirical evidence challenging two classical views in semantics, and we develop a novel account of counterfactuals that combines ideas from inquisitive semantics and causal reasoning. First, we show that two truth-conditionally equivalent clauses can make different semantic contributions when embedded in a counterfactual antecedent. Assuming compositionality, this means that the meaning of these clauses is not fully determined by their truth conditions. This finding has a clear explanation in inquisitive semantics: truth-conditionally equivalent clauses may be associated with different propositional alternatives, each of which counts as a separate counterfactual assumption. Second, we show that our results contradict the common idea that the interpretation of a counterfactual involves minimizing change with respect to the actual state of affairs. Building on techniques from causal reasoning, we propose to replace the idea of minimal change by a distinction between foreground and background for a given counterfactual assumption: the background is held fixed in the counterfactual situation, while the foreground can be varied without any minimality constraint.

Keywords: counterfactuals, disjunctive antecedents, minimal change semantics, inquisitive semantics, web survey, causal reasoning

1 Introduction

Imagine a long hallway with a light in the middle and with two switches, one at each end. One switch is called switch A and the other one is called switch B. As the following wiring diagram shows (see Figure 1), the light is on whenever both switches are in the same position (both up or both down); otherwise, the light is off. Right now, switch A and switch B are both up, and the light is on. But things could be different…

Which of the following counterfactual sentences are true in this scenario?

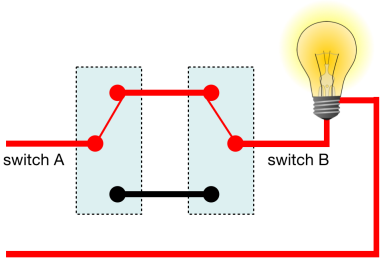


Figure 1: A multiway switch

(1) a. If switch A was down, the light would be off. $\overline{A} > \textsc{off}$
b. If switch B was down, the light would be off. $\overline{B} > \textsc{off}$
c. If switch A or switch B was down, the light would be off. $\overline{A} \vee \overline{B} > \textsc{off}$
d. If switch A and switch B were not both up, the light would be off. $\neg(A \wedge B) > \textsc{off}$
e. If switch A and switch B were not both up, the light would be on. $\neg(A \wedge B) > \textsc{on}$

This simple empirical question bears on two fundamental issues in semantics, namely, (i) the nature of sentence meaning, and (ii) the interpretation of counterfactual conditionals. Motivated by experimental results, this paper challenges classical views on these two issues and develops a novel account of counterfactuals, which combines ideas from inquisitive semantics and causal reasoning.

For the sake of readability, throughout the paper we will refer to the sentences in (1) by the mnemonic labels that appear next to them. These labels are based on the following conventions: we write A as an abbreviation for *switch A is up*, $\overline{A}$ for *switch A is down*, and similarly for switch B. We use the standard logical notations $\neg, \wedge, \vee$ for negation, conjunction, and disjunction, and $>$ for the counterfactual conditional construction. In later sections, we will assume that these representations correspond to the logical forms of these sentences, at a suitable level of abstraction.

1.1 The nature of sentence meaning

The first issue we investigate is the relation between sentence meaning and truth conditions. There are two views on this issue. The textbook view is that truth conditions completely determine meaning: "To know the meaning of a sentence is to know its truth conditions" (Heim & Kratzer 1998: p. 1). In the standard intensional semantics framework, this view is implemented by representing the meaning of a sentence as a set of possible worlds—the set of those worlds in which the sentence is true.

An alternative view is that the meaning of a sentence carries some extra structure beyond what is needed to capture its truth conditions, and that the notion of sentential meaning is therefore more fine-grained than what is provided by sets of possible worlds. Here, we focus on a particular framework instantiating this view, inquisitive

semantics (Ciardelli, Groenendijk & Roelofsen 2013).[1] In inquisitive semantics, the meaning of a sentence is not represented as a set of worlds, but as a set of such sets, referred to as the *alternatives* for the sentence.[2] By representing sentence meaning as a set of alternatives, inquisitive semantics captures the intuition that a single sentence can evoke several possibilities. A sentence is true at a world w if w belongs to some alternative for the sentence. Consequently, in inquisitive semantics, the meaning of a sentence determines its truth conditions, but the converse is not the case: two sentences may have the same truth conditions but be associated with different alternatives, and thus with different meanings.

The difference between these two views can be illustrated with a simple example. Taking for granted that the switches of our scenario assume only two positions, *up* and *down*, the following sentences have the same truth conditions.

(2) a. Switch A or switch B is down. $\overline{A}\vee\overline{B}$
b. Switch A and switch B are not both up. $\neg(A\wedge B)$

Whenever $\overline{A}\vee\overline{B}$ is true, $\neg(A\wedge B)$ is true as well, and vice versa (we will provide experimental evidence for this in Section 2.3.1). In fact, given the assumption that a switch is down whenever it is not up, the truth-conditional equivalence between $\overline{A}\vee\overline{B}$ and $\neg(A\wedge B)$ can be regarded as an instance of de Morgan's law $\neg A\vee\neg B \equiv \neg(A\wedge B)$, which is valid in classical logic—the logic arising from truth-conditional semantics. On the textbook view, sentences $\overline{A}\vee\overline{B}$ and $\neg(A\wedge B)$ therefore have the same meaning.

By contrast, on the view that meaning is not completely determined by truth conditions, $\overline{A}\vee\overline{B}$ and $\neg(A\wedge B)$ may well differ in meaning. This is in fact the case in inquisitive semantics, where $\overline{A}\vee\overline{B}$ is associated with two distinct alternatives, corresponding to the two disjuncts, while $\neg(A\wedge B)$ is associated with a unique alternative.

Under the assumption of compositionality, the two approaches lead to different expectations about what should happen when sentences like $\overline{A}\vee\overline{B}$ and $\neg(A\wedge B)$ are embedded as constituents in a larger sentence. On the textbook view, since $\overline{A}\vee\overline{B}$ and $\neg(A\wedge B)$ have the same meaning, we expect them to make exactly the same semantic contribution to the sentences they are part of. By contrast, in a framework such as inquisitive semantics, where $\overline{A}\vee\overline{B}$ and $\neg(A\wedge B)$ have different meanings, we expect that these sentences could make different semantic contributions, and that this may sometimes lead to different truth conditions for the resulting sentences.

For instance, consider the counterfactuals $\overline{A}\vee\overline{B} > \textsc{off}$ and $\neg(A\wedge B) > \textsc{off}$, which embed $\overline{A}\vee\overline{B}$ and $\neg(A\wedge B)$ as antecedents. If $\overline{A}\vee\overline{B} > \textsc{off}$ and $\neg(A\wedge B) > \textsc{off}$ turn out to have different truth conditions, this is incompatible with compositionality combined with the textbook view on meaning, and it provides an empirical argument for distinguishing the meanings assigned to $\overline{A}\vee\overline{B}$ and $\neg(A\wedge B)$.

[1] Other semantic frameworks such as alternative semantics (Hamblin 1973, Kratzer & Shimoyama 2002, Alonso-Ovalle 2006) and truth-maker semantics (Fine 2012a) are similar to inquisitive semantics in this respect. However, some specific differences between these frameworks also matter for our concerns, as we discuss in Section 6.2. Dynamic theories of meaning (Kamp 1981, Heim 1982, Groenendijk & Stokhof 1990) also take the view that meaning is not completely determined by truth conditions, but they are primarily concerned with discourse phenomena and anaphora, which we set aside here.

[2] The semantic notion of alternatives at stake here should not be confused with the notion of focus alternatives (Rooth 1985), or with the notion of scalar alternatives (e.g., Horn 1972).

To investigate whether $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF indeed come apart in their truth values, we conducted a truth-value judgment experiment based on the context described above. Our experimental results show that while a vast majority of participants judged $\overline{A} \vee \overline{B} >$ OFF true in the given scenario, few participants judged $\neg(A \wedge B) >$ OFF true.

In addition to offering empirical evidence in favor of a fine-grained view of meaning that teases apart $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$, in this paper we also develop a formal theory that explains how the truth-conditional difference between the counterfactuals $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF arises from the non-truth-conditional difference between their antecedents $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$. Our theory combines an inquisitive semantic account of the propositional connectives with a proposal by Alonso-Ovalle (2006, 2009) for counterfactuals (see also van Rooij 2006). According to this proposal, a counterfactual antecedent does not always provide a unique assumption: when it is associated with multiple alternatives, as in the case of $\overline{A} \vee \overline{B}$, each of these alternatives introduces a separate assumption into the compositional process. Since $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$ are associated with different alternatives, the consequents of $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF are assessed in different hypothetical situations, resulting in different truth conditions.

1.2 The interpretation of counterfactual conditionals

Aside from providing a probe into the nature of sentence meaning, our empirical observations also bear on a fundamental issue concerning the interpretation of counterfactuals. This issue consists in determining which hypothetical situations should be considered in order to assess the truth of a counterfactual, and which ones should be set aside.

The standard view is that the situations that should be considered are those where the antecedent is true, and which otherwise differ minimally from the actual situation. We refer to this as the *minimal change requirement*. This view is at the heart of the most influential theories of counterfactuals (Stalnaker 1968, Lewis 1973, Kratzer 1981a, Pearl 2000). One of the main goals of our paper is to show that the minimal change requirement leads to incorrect predictions concerning the interpretation of counterfactuals with complex antecedents, and to present an alternative view.

The minimal change requirement was motivated by counterfactuals with simple antecedents such as this classical example:

(3) If this match was struck, it would light.

In judging this sentence, we allow certain background facts to carry over from the actual world to the hypothetical situations we consider. For example, we assume that the match was not soaked in water overnight, that there is oxygen in the air, etc. The purpose of the minimal-change requirement is to prevent us from introducing any gratuitous changes to such facts.

In the canonical account of counterfactuals, ordering semantics (Lewis 1973), the minimal change requirement is implemented as follows. Counterfactuals are interpreted by means of a relation of comparative similarity to the world of evaluation. This relation is assumed to be a weak total order on possible worlds (that is, a total

order that allows for ties). Intuitively, a world w' counts as more similar than w'' to the world of evaluation w just in case getting from w to w' involves a smaller amount of change than getting from w to w''. Glossing over details that do not matter for our argument, the main idea of ordering semantics is that a counterfactual $\varphi > \psi$ is true at a world w in case ψ is true at each of the worlds which are most similar to w among the worlds where φ is true.

Soon after ordering semantics was proposed, there was a debate about whether the minimal change requirement properly characterizes the interpretation of counterfactuals. This debate is exemplified by Fine (1975) and Lewis (1979), who considered the following sentence.

(4) If Nixon had pressed the button there would have been a nuclear holocaust.

This counterfactual is true, or can be imagined to be so. However, Fine argues, under the minimal change requirement it would be predicted to be false. For consider a possible world in which Nixon has pressed the button. Any world in which a nuclear holocaust results will differ much more sharply from the actual world than a world in which a simple wire malfunction occurs, preventing nuclear war. Hence, the most similar worlds in which Nixon presses the button are ones in which no nuclear holocaust takes place. As a consequence, (4) is predicted to be false.

This argument, however, relies in a crucial way on certain assumptions about the notion of similarity: a world where a wire malfunction occurs must count as more similar to the real world than one where nuclear war takes place. While this is in line with the intuitive notion of similarity, Lewis (1979) argued that this is not the notion that matters for the evaluation of counterfactuals; instead, he proposed a principled theory that delivers a similarity ordering on which (4) is true. Thus, arguments such as Fine's do not threaten the minimal change requirement in any obvious way, and the question whether this requirement properly characterizes the interpretation of counterfactuals remains open.[3]

In this paper we test the predictions of the minimal change requirement in a novel way. Crucially, we do not need to rely on any pre-defined assumption about similarity. In a given context, we can use truth value intuitions about some counterfactuals to infer what the relevant similarity ordering must be like for these intuitions to be accounted for; we can then use our findings about similarity to predict the truth value of another counterfactual, and check whether this prediction is empirically correct.[4]

Consider our scenario above. For conciseness, let us refer to a world where a sentence φ is true as a φ-world. Suppose that the counterfactuals $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ are true in the given situation. This means that the relevant similarity ordering must be such that all the most similar $\overline{A}$-worlds are Off-worlds, and analogously for the most similar $\overline{B}$-worlds. It turns out that this is sufficient to make a prediction about the truth of $\neg(A \wedge B) > \textsc{off}$. For consider a most similar $\neg(A \wedge B)$-world, i.e., a most

[3] In effect, Lewis argued that a similarity ordering that predicts (4) true is not as far-fetched as it may seem. For the wire to spontaneously become faulty, a "miracle" (a violation of the laws of nature) would be necessary, and that would count as a larger change than even nuclear war. This theory in turn led to further proposals and debates about the nature of the similarity ordering; for an overview, see Menzies (2014).

[4] This way of arguing is in line with a methodological suggestion by Lewis (1979: p. 466f.).

similar world where the switches are not both up. This must be either a most similar $\overline{A}$-world or a most similar $\overline{B}$-world. In either case, the light must then be off in this world. Consequently, $\neg(A \wedge B) > \textsc{off}$ is predicted to be true.[5]

Thus, regardless of what notion of world similarity we assume, ordering semantics predicts that if $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ are true, then so is $\neg(A \wedge B) > \textsc{off}$. In other words, the entailment $\overline{A} > \textsc{off}$, $\overline{B} > \textsc{off} \models \neg(A \wedge B) > \textsc{off}$ is logically valid in ordering semantics. Although we formulated this argument in the context of ordering semantics, an analogous conclusion can be reached in other frameworks that incorporate the minimal change requirement, such as premise semantics (Kratzer 1981a,b). We come back to this in Section 6.3.

A truth-value judgment task including $\overline{A} > \textsc{off}$, $\overline{B} > \textsc{off}$, and $\neg(A \wedge B) > \textsc{off}$ allowed us to test whether this prediction of the minimal-change requirement is borne out. Our findings contradict this prediction: while a vast majority of participants judged both $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ true in the given scenario, few judged $\neg(A \wedge B) > \textsc{off}$ true.

Aside from presenting empirical evidence against the minimal change requirement, in this paper we provide a conceptual explanation of our findings and a corresponding formal account. We replace the idea of minimal change by a distinction between aspects of the actual situation that are at stake when making a counterfactual assumption and aspects that are regarded as background. While background facts are held fixed in making the assumption, aspects that are at stake can be varied without any minimality constraints.

The intuitive idea is that when making the counterfactual assumption that A is down, the position of B is not at stake and can be regarded as background; since background facts are held fixed, in the counterfactual scenario switch A is down, but switch B is still up; by the laws of the circuit, this implies that the light is off. This explains why $\overline{A} > \textsc{off}$ is judged true. The situation is analogous for $\overline{B} > \textsc{off}$. However, when making the assumption that A and B are not both up, the positions of both switches are now at stake, and neither can be regarded as background; thus, in the counterfactual scenario, nothing about the actual situation is retained, and all we can assume is that A and B are not both up. Since this does not allow us to reach any definite conclusion about the state of the light, $\neg(A \wedge B) > \textsc{off}$ is not judged true.

Building on ideas from the literature on causal reasoning (Pearl 2000), we develop a formal theory of counterfactuals that embodies these intuitions, and we show that, in combination with inquisitive semantics, this theory accounts for our findings.

1.3 Structure of the paper

The paper is organized as follows. In Section 2 we present the details of our experiment. We argue that our experimental results raise two problems, one for the view that meaning can be equated with truth conditions, and the other for the minimal change requirement. Section 3 shows how the first problem can be solved by lifting an account of conditionals to inquisitive semantics. Section 4 shows how the second problem can be solved by replacing the minimal change requirement by the idea of

[5] The reader who worries about the distinction between a switch being down and it not being up should just replace the word *down* by *not up* in our sentences. As we will see in Section 2.5.2, this does not affect our empirical findings.

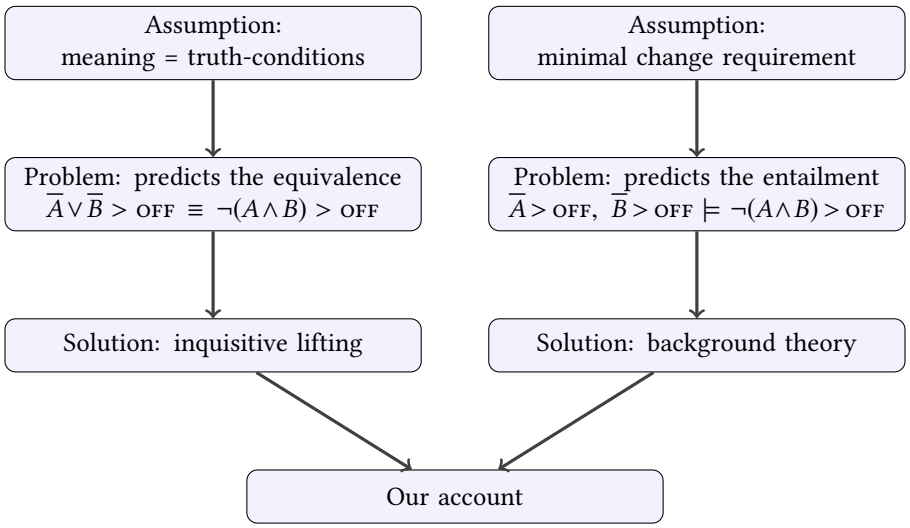


Figure 2: The paper at a glance.

a factual background. Section 5 extends the empirical scope of our observations to other classes of conditionals, and discusses additional patterns in our data. Section 6 discusses the connections between our proposal and related work on counterfactuals and modals. Section 7 sums up and concludes. A visual outline of the challenges we discuss and of the solutions we propose is given in Figure 2.

2 Experiment

2.1 Hypotheses and predictions

Throughout this paper, we assume the principle of semantic compositionality for natural language; that is, the meaning of a complex expression is completely determined by the meaning of its constituents and the way they are combined. This implies that the meaning of a complex expression does not change when one of its constituents φ is replaced by another expression φ' with the same meaning.

The two questions we seek to answer are whether the truth conditions of a sentential clause completely determine its meaning, and whether the interpretation of counterfactuals with complex antecedents challenges the minimal change requirement.

For the first question, our experiment took advantage of the truth-conditional equivalence between $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$. Under the hypothesis that truth conditions completely determine meaning, $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$ have the same meaning. Consequently, given compositionality, the two counterfactuals $\overline{A} \vee \overline{B} > \text{OFF}$ and $\neg(A \wedge B) > \text{OFF}$, which embed respectively $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$ but are otherwise identical, should have the same meaning as well, and should be judged in the same way in the given situation.

By contrast, under the hypothesis that sentence meanings are not fully determined by their truth conditions, $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$ may well have different meanings. If so, these clauses could make a different contribution when embedded in a conditional antecedent, which may result in $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF having different truth values in the given situation—something that would be reflected by native speakers' truth value intuitions.

To answer the second question, we tested native speakers' judgments of $\overline{A} >$ OFF, and $\overline{B} >$ OFF, and $\neg(A \wedge B) >$ OFF. As we argued in Section 1.2, the minimal change requirement predicts that if $\overline{A} >$ OFF and $\overline{B} >$ OFF are both judged true then $\neg(A \wedge B) >$ OFF should be judged true as well. Thus, if $\overline{A} >$ OFF and $\overline{B} >$ OFF but not $\neg(A \wedge B) >$ OFF are judged true, the minimal change requirement is obviously challenged.

2.2 Experiment design and methods

Our experiment included three parts: (i) two pretests (Section 2.3), (ii) the main experiment (Section 2.4), and (iii) three post-hoc tests (Section 2.5). Pretest I confirmed the truth-conditional equivalence between $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$ for native speakers of English, and Pretest II confirmed that the critical sentences used in our main experiment, namely, $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF, are natural to native speakers to the same degree. In the main experiment, we elicited native speakers' truth value judgments for the counterfactuals in (1). We used the three post-hoc tests to rule out some alternative accounts for the findings of our main experiment.

We implemented all these experiments and tests as web surveys using TurkTools (Erlewine & Kotek 2016), which relies on the online labor market platform Amazon Mechanical Turk (http://www.mturk.com). Participants were all required to be located in the United States and have a Mechanical Turk approval rate (an indication of reliability) of at least 95%.

In all tests, each participant was asked to judge two sentences: one target and one filler sentence. For half of the participants, the target preceded the filler, while for the other half, the order of presentation was reversed. Our fillers were all uncontroversial in terms of naturalness or truth value, and thus the response to them was an indication showing whether participants paid enough attention to stimuli.

In the main experiment, Pretest I, and the three post-hoc tests, participants were shown a pictorial context[6] along with a short descriptive text and were asked to judge whether what the sentences say about the picture is 'true', 'false' or 'indeterminate'.

In Pretest II, there was no pictorial context or descriptive text. Participants were asked to judge whether the sentences sound natural on a 7-point scale, where 1 stands for "totally unnatural" and 7 for "perfectly natural".

Before the presentation of our stimuli, we gave examples illustrating the truth value or naturalness judgment task. At the end of the survey, we asked participants whether they were native speakers of English, whether they spoke British or American English or another dialect, and whether they had any comments for us (few did). We stated that their answers to these questions would not affect the payment.

[6]Our figures are adapted from multiway switches © Cburnett (https://en.wikipedia.org/wiki/Multiway_switching#/media/File:3-way_switches_position_2.svg) CC BY-SA 3.0.

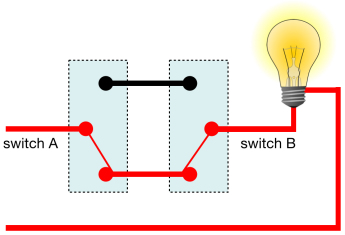


Figure 3: Pretest I. Switch A and B are both down, and the light is on.

For the truth value judgment task,we paid each participant $0.10. For the naturalness judgment task,we paid each participant $0.02. We used participants' responses to demographic questions and filler sentences to filter data: responses from those who did not self-identify as native speakers of American English or who failed to judge the filler sentence correctly were ruled out from further analyses. If someone took part in our study more than once, only their first response was included in data analysis. In all tests, incorrect responses to filler items accounted for the vast majority of rejected data.

All our experimental materials, raw data, scripts for data processing and analysis, as well as a detailed summary of results are available in the supplementary material of this paper.

2.3 Two pretests

2.3.1 Pretest I

Materials The goal of Pretest I was to confirm that $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$ have identical truth conditions. Since these sentences are both undoubtedly true when exactly one of the two switches is down, and false when both switches are up, we only elicited truth value judgments of these sentences in a scenario where both switches are down.

To this end, we used the pictorial context in Figure 3 and asked participants to provide truth value judgments for $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$. We also included the sentence *Switch A is up* as a filler item, and we discarded data from participants who failed to judge it false in this context.

Results We collected data from 330 non-repetitive participants who are native speakers of American English and rejected 16.67% of the responses. As shown in Table 1, each sentence was judged true by over 80% of participants. As expected, the results for both sentences did not differ significantly ($\chi^2(2, N = 275) = 5.23$, $p = 0.07$). We conclude that $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$ are indeed truth-conditionally equivalent.

Table 1: Results of Pretest I

Sentence	Number	True	(%)	False	(%)	Indeterminate	(%)
$\overline{A}\vee\overline{B}$	145	118	81.38%	23	15.86%	4	2.76%
$\neg(A\wedge B)$	130	118	90.77%	11	8.46%	1	0.77%

Table 2: Results of Pretest II

Sentence	Label	Number	Mean rating	Standard deviation
$\overline{A}\vee\overline{B} > \textsc{off}$	$\overline{A}\vee\overline{B} > \textsc{off}$	73	5.07	1.63
$\neg(A\wedge B) > \textsc{off}$	$\neg(A\wedge B) > \textsc{off}$	55	5.16	1.76

2.3.2 Pretest II

Materials We assume that counterfactual sentences with simple antecedents (e.g., *if switch A was down, the light would be off*) are natural. Here in Pretest II, we aimed to verify that the two counterfactual sentences with complex antecedents, $\overline{A}\vee\overline{B} > \textsc{off}$ and $\neg(A\wedge B) > \textsc{off}$, sound equally natural to native speakers. We used the sentence *If I were in the hallway, I would turn the light off* as the filler item, and we excluded data from participants who judged the filler lower than 5 on a 7-point scale.

Results As shown in Table 2, both sentences were judged acceptable at comparable levels: the t-test comparing the scores of these two sentences showed no significant difference ($p = 0.37$). Thus, any potential differences between the truth value judgments of these two sentences are unlikely to be attributable to one of the sentences being less natural than the other.

2.4 Main experiment

In our main experiment, we presented the context described in the introduction, and we asked participants to give truth value judgments for one of the five counterfactual sentences in (1).

Materials Our context consisted of the descriptive text at the outset of the paper, repeated below, and of Figure 1, repeated here as Figure 4.[7]

(5) Imagine a long hallway with a light in the middle and with two switches, one at each end. One switch is called switch A and the

[7] The two-switches scenario was originally introduced by Lifschitz (1990) in the context of causal reasoning. Within the literature on counterfactuals, it was first discussed in Schulz (2007), as a counterexample to the theory in Veltman (2005). That discussion is not directly related to our main concerns here. The specific text in (5) is our own, and to the best of our knowledge, our paper is the first to discuss the two-switches scenario in connection with complex antecedents.

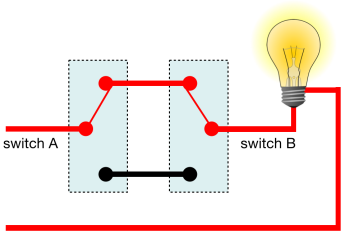


Figure 4: Main experiment. Switch A and B are both up, and the light is on.

> other one is called switch B. As this wiring diagram shows, the light is on whenever both switches are in the same position (both up or both down); otherwise, the light is off. Right now, switch A and switch B are both up, and the light is on. But things could be different …

Our five target sentences, repeated from (1), are shown in (6) The labels that appear next to them were not shown to participants.

(6)
a. If switch A was down, the light would be off. $\overline{A} > \textsc{off}$
b. If switch B was down, the light would be off. $\overline{B} > \textsc{off}$
c. If switch A or switch B was down, the light would be off. $\overline{A} \vee \overline{B} > \textsc{off}$
d. If switch A and switch B were not both up, the light would be off. $\neg(A \wedge B) > \textsc{off}$
e. If switch A and switch B were not both up, the light would be on. $\neg(A \wedge B) > \textsc{on}$

Our filler sentence is shown in (7). We ruled out data from those participants who failed to judged it false in the given context.

(7) If switch A and switch B were both down, the light would be off.

Results We collected data from 2299 non-repetitive participants who are native speakers of American English and rejected 38.02% of the responses. The remaining 1425 responses are summarized in Table 3.

Differences across all five sentences were highly significant ($\chi^2(8, N = 1425) = 383.36, p < 0.0001$). Our results fall naturally into two blocks, as indicated by the dashed line in Table 3.[8] The first block consists of $\overline{A} > \textsc{off}$, $\overline{B} > \textsc{off}$, and $\overline{A} \vee \overline{B} > \textsc{off}$, which were all judged true by a wide majority. In the second block, $\neg(A \wedge B) > \textsc{off}$ and $\neg(A \wedge B) > \textsc{on}$ were generally judged false or indeterminate. The frequency difference between $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$ is highly significant: $\chi^2(2, N = 734) =$

[8] This observation is confirmed by statistical analysis. All pairwise chi-square tests between sentences across blocks are highly significant ($p < 0.0001$); pairwise comparisons within each block are not ($\overline{A} > \textsc{off}$ vs. $\overline{B} > \textsc{off}$: $\chi^2(2, N = 491) = 0.90$; $\overline{A} > \textsc{off}$ vs. $\overline{A} \vee \overline{B} > \textsc{off}$: $\chi^2(2, N = 618) = 0.28$; $\overline{B} > \textsc{off}$ vs. $\overline{A} \vee \overline{B} > \textsc{off}$: $\chi^2(2, N = 597) = 0.37$; $\neg(A \wedge B) > \textsc{off}$ vs. $\neg(A \wedge B) > \textsc{on}$: $\chi^2(2, N = 572) = 0.38$).

Table 3: Results of the main experiment

Sentence	Number	True	(%)	False	(%)	Indet.	(%)
$\overline{A} >$ OFF	256	169	66.02%	6	2.34%	81	31.64%
$\overline{B} >$ OFF	235	153	65.11%	7	2.98%	75	31.91%
$\overline{A} \vee \overline{B} >$ OFF	362	251	69.33%	14	3.87%	97	26.80%
$\neg(A \wedge B) >$ OFF	372	82	22.04%	136	36.56%	154	41.40%
$\neg(A \wedge B) >$ ON	200	43	21.50%	63	31.50%	94	47.00%

$197.84, p < 0.0001$. Differences within each block were not significant (first block: $\chi^2(4, N = 853) = 3.33, p = 0.5042$; second block: $\chi^2(2, N = 572) = 1.92, p = 0.3829$).

Crucially, our results show that $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF were judged differently, indicating that these two counterfactuals have different truth conditions. Moreover, $\neg(A \wedge B) >$ OFF was judged false by most participants, while $\overline{A} >$ OFF and $\overline{B} >$ OFF were judged true, contrary to the predictions of the minimal change requirement.

2.5 Three post-hoc tests

The findings of our main experiment suggest that the clauses $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$ differ in meaning, contradicting the view that meaning can be equated with truth conditions. Moreover, they suggest that in our context, $\overline{A} >$ OFF and $\overline{B} >$ OFF are true, while $\neg(A \wedge B) >$ OFF is false, contrary to the predictions of the minimal change requirement.

To solidify these conclusions, we ran three post-hoc tests that rule out some potential alternative explanations for the drop in 'true' judgments from the first three sentences, $\overline{A} >$ OFF, $\overline{B} >$ OFF and $\overline{A} \vee \overline{B} >$ OFF, to the fourth, $\neg(A \wedge B) >$ OFF.

2.5.1 Post-hoc test I: the light is on only if both switches are up

Materials Post-hoc test I aimed to test whether the judgments reported in our main experiment might be due to context-independent factors such as differences in complexity or processing load. To this end, we replaced the pictorial context by the one shown in Figure 5, in which the light is on only if both switches are up, and we replaced the third sentence in our descriptive text by the sentence in (8):

(8) As the following wiring diagram shows, the light is on whenever both switches are up; otherwise, the light is off.

If the judgments reported in our main experiment are mainly due to context-independent factors, we should observe exactly the same difference in this post-hoc test. Alternatively, if it indeed tracks actual differences in truth conditions, then in this new context, we expect that the result pattern for the five counterfactual sentences might change.

We used the filler *If switch A and switch B were both down, the light would be on*, and we rejected data from participants who failed to judge the filler false.

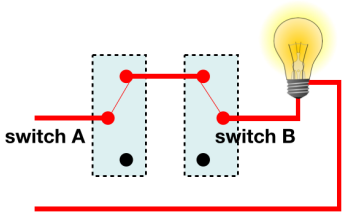


Figure 5: Post-hoc test I. There is no wire between the two "down" positions.

Table 4: Results of Post-hoc test I

Sentence	Number	True	(%)	False	(%)	Indet.	(%)
$\overline{A} >$ OFF	52	41	78.85%	5	9.61%	6	11.54%
$\overline{B} >$ OFF	68	60	88.24%	5	7.35%	3	4.41%
$\overline{A} \vee \overline{B} >$ OFF	110	104	94.55%	1	0.91%	5	4.54%
$\neg(A \wedge B) >$ OFF	116	99	85.34%	9	7.76%	8	6.90%
$\neg(A \wedge B) >$ ON	103	19	18.45%	79	76.70%	5	4.85%

Results We collected data from 553 non-repetitive participants who are native speakers of American English and rejected 18.81% of the responses. The remaining 449 responses are summarized in Table 4.

This time, there were no significant differences among the truth value judgments of the first four sentences ($\chi^2(6, N = 346) = 11.26, p = 0.08$). Moreover, for both $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF, most ($> 85\%$) participants judged them to be true in this context. These results suggest that the difference in truth value judgments between $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF that we observed in our main experiment is unlikely to be due to context-independent factors. Similarly, this time $\overline{A} >$ OFF, $\overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF were all judged true by a large majority, suggesting that in our main test, the judgment difference between $\overline{A} >$ OFF/$\overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF is not due to context-independent factors.

2.5.2 Post-hoc test II: replacing *down* by *not up*

Materials Post-hoc test II was designed to test whether the presence or absence of explicit negation affects the result pattern of the main experiment. To this end, we replaced the word *down* by *not up* in the target sentences that used it. We did not replace *down* by *not up* in the filler sentence.

Results For $\neg A >$ OFF, $\neg B >$ OFF, and $\neg A \vee \neg B >$ OFF, we collected data from 561 non-repetitive participants who are native speakers of American English and rejected 71.66% of the responses.[9] The remaining 159 responses are summarized in Table 5

[9]In both Post-hoc tests II and III, a large proportion of data were rejected due to participants being incorrect in answering the filler item (71.66% for Post-hoc test II, and 67.27% for Post-hoc test III). This is

Table 5: Results of Post-hoc test II. The last two lines are repeated from Table 3.

Sentence	Number	True	(%)	False	(%)	Indet.	(%)
$\neg A >$ OFF	36	27	75.00%	1	2.78%	8	22.22%
$\neg B >$ OFF	43	28	65.12%	7	16.28%	8	18.60%
$\neg A \vee \neg B >$ OFF	80	48	60.00%	16	20.00%	16	20.00%
$\neg(A \wedge B) >$ OFF	372	82	22.04%	136	36.56%	154	41.40%
$\neg(A \wedge B) >$ ON	200	43	21.50%	63	31.50%	94	47.00%

along with the results of $\neg(A \wedge B) >$ OFF and $\neg(A \wedge B) >$ ON from the main experiment.

Table 5 shows that substituting *not up* for *down* did not change the pattern in the observed results: differences across all five sentences were highly significant ($\chi^2(8, N = 743) = 129.26, p < 0.0001$). The results shown in Table 5 also fall naturally into two blocks, as indicated by the dashed line. Sentences of the first block were all judged true by a majority, and differences within the first block were not significant ($\chi^2(4, N = 159) = 5.93, p = 0.2044$). The difference between $\overline{A} \vee \overline{B} >$ OFF in this test and $\neg(A \wedge B) >$ OFF in the main experiment is still significant: $\chi^2(2, N = 452) = 46.37, p < 0.0001$. Therefore, we can exclude the presence or absence of the word *not* as a potential confounding factor. This also confirms our background assumption that few participants, if any, would consider the possibility that a switch might be in an intermediate position (that is, neither up nor down).

2.5.3 Post-hoc test III: replacing *was* by *were*

Materials Post-hoc test III was designed to rule out the possibility that the choice of auxiliary affected the truth value judgments we found in our main experiment. To this end, we replaced the word *was* by *were* in the target sentences that used it ($\overline{A} >$ OFF, $\overline{B} >$ OFF, and $\overline{A} \vee \overline{B} >$ OFF).

Results We collected data from 556 non-repetitive participants who are native speakers of American English and rejected 67.27% of the responses. The remaining 182 responses are summarized in Table 6.

Overall, the results of Post-hoc test III yielded the same pattern as in the main experiment. Each of the sentences in this test was judged true by most (> 70%) of the participants. Moreover, the difference between $\overline{A} \vee \overline{B} >$ OFF in this test and $\neg(A \wedge B) >$ OFF in the main experiment is still significant: $\chi^2(2, N = 455) = 83.89, p < 0.0001$.

mysterious, and we only have a conjecture here: using *not up* instead of *down* and using *were* instead of *was* degraded the naturalness of sentences and thus made participants confused and their truth value judgments less reliable. In a separate naturalness test, we used these two factors to construct four sentences-to-test (*If switch A was/were down/not up, the light would be off*) and conducted a 2 by 2 ANOVA, which indeed revealed that *down*-sentences (N = 63, Mean = 5.41, SD = 1.47) were rated significantly more natural than *not-up*-sentences (N = 63, Mean = 4.33, SD = 1.69) ($F(1, 122) = 14.56$, $p < 0.001$), and *was*-sentences (N = 63, Mean = 5.03, SD = 1.61) were also rated more natural than *were*-sentences (N = 63, Mean = 4.71, SD = 1.73) numerically, though this difference was not significant ($F(1, 122) = 1.26$, $p = 0.26$).

Table 6: Results of Post-hoc test III

Sentence	Number	True	(%)	False	(%)	Indet.	(%)
$\overline{A} >$ OFF	57	46	80.70%	0	0%	11	19.30%
$\overline{B} >$ OFF	42	35	83.33%	2	4.76%	5	11.90%
$\overline{A} \vee \overline{B} >$ OFF	83	61	73.49%	13	15.66%	9	10.84%

Therefore, we can exclude the choice of auxiliary as a potential factor affecting our main finding.[10]

2.6 Discussion and conclusions

2.6.1 Summary of experimental findings

As shown in the results of our main experiment (Table 3), $\overline{A} >$ OFF, $\overline{B} >$ OFF, and $\overline{A} \vee \overline{B} >$ OFF were generally judged true. Given the way the switches are wired, this suggests that most participants interpreted $\overline{A} >$ OFF and $\overline{B} >$ OFF by considering what would be the case if just the switch in question was toggled, leaving the other one in place. Similarly, it seems that most participants interpreted $\overline{A} \vee \overline{B} >$ OFF by considering one switch at a time, while ignoring the option that both switches might be toggled simultaneously.[11]

As for $\neg(A \wedge B) >$ OFF and $\neg(A \wedge B) >$ ON, most participants judged them indeterminate or false. This suggests that the predominant strategy for these sentences is to consider all three possibilities: only switch A is toggled; only switch B is toggled; both switches are toggled. These possibilities do not all agree on the state of the light, leading to the lack of 'true' judgments.

2.6.2 Ruling out alternative accounts for our findings

Nute (1975: p. 775) concludes from an example whose logic is similar to that of our scenario that some instances of natural language *or* in counterfactual antecedents are interpreted as exclusive rather than inclusive disjunction. The possibility that *or* is lexically ambiguous between inclusive and exclusive meanings is generally seen as problematic (Aloni 2016). However, some previous studies (e.g. Fox 2007, Spector 2007) have suggested that a silent exhaustivity operator *Exh* might strengthen the meaning of natural language *or* and be responsible for what appears to be an exclusive

[10] This time, the comparison among the three sentences $\overline{A} >$ OFF, $\overline{B} >$ OFF and $\overline{A} \vee \overline{B} >$ OFF showed a significant difference: $\chi^2(4, N = 182) = 13.18, p = 0.01$. While have no explanation for this fact, this seems orthogonal to our main concern in this experiment, which was to show that the presence of the auxiliary *were* cannot be responsible for the drop in 'true' judgments that we observe in our main experiment between $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF.

[11] The filler sentence queried that option; by discarding data from participants who judged it incorrectly, we guarded against the possibility that participants were unaware of the fact that the light remains on when both switches are toggled.

interpretation in certain environments. One may wonder whether such a strengthening is responsible for the observed difference between $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF. However, Pretest I has shown that $\overline{A} \vee \overline{B}$, the main clause corresponding to the antecedent of $\overline{A} \vee \overline{B} >$ OFF, is generally judged true even when both switches are down. This means that exhaustive strengthening generally does not take place in $\overline{A} \vee \overline{B}$. We know of no evidence that strengthening happens in counterfactual antecedents more often than in main clauses.[12] Therefore, it seems unlikely that the observed difference between $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF can be attributed to the fact that a silent exhaustivity operator strengthens the antecedent of $\overline{A} \vee \overline{B} >$ OFF.

Since Pretest II has shown that $\neg(A \wedge B) >$ OFF and $\overline{A} \vee \overline{B} >$ OFF are judged equally natural, the drop in 'true' judgments between these two sentences cannot be due to differences in sentence naturalness. Relatedly, Post-hoc tests II and III showed that the drop in 'true' judgments between these two sentences cannot be attributed to the use of *were* or explicit negation in $\neg(A \wedge B) >$ OFF, either.

Finally, it is conceivable that in $\neg(A \wedge B) >$ OFF, the string *not both up* might have been interpreted as *both not up*, either through misreading or as a result of interpreting *up* as focused (Rooth 1996). However, we separately tested the sentence *Switch A and switch B are not both up* in a pictorial context that shows switch A up and switch B down, and 76.9% of 290 participants judged it true, showing that interpreting *not both up* as *both not up* is at best unlikely. Moreover, if participants really interpreted *not both up* as *both down*, we would expect a spike in 'true' judgments for $\neg(A \wedge B) >$ ON; but only 21.5% of participants judged this sentence true.

2.6.3 Conclusion

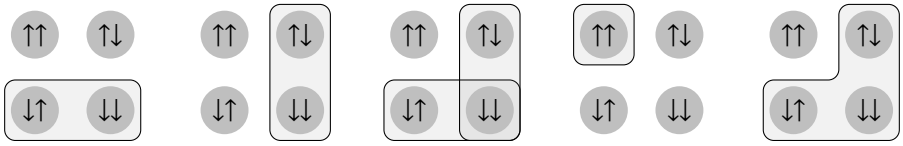
Having excluded various confounds, we conclude that the differences we observed in native speakers' judgments on our sentences track actual differences in the truth values of these sentences: in our context, $\overline{A} >$ OFF, $\overline{B} >$ OFF and $\overline{A} \vee \overline{B} >$ OFF are true, while $\neg(A \wedge B) >$ OFF and $\neg(A \wedge B) >$ ON are not.

Recall that the two questions we seek to answer are whether the truth conditions of a sentential clause completely determine its meaning, and whether the interpretation of counterfactuals with complex antecedents challenges the minimal change requirement.

With respect to the first question, we take our results to show that $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF have different meanings. By compositionality, their antecedents, corresponding to $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$, must then have different meanings as well. However, these antecedents have the same truth conditions, as confirmed by Pretest I. Hence, it is possible for two sentential clauses to have the same truth conditions and different meanings—which shows that meaning is not completely determined by truth conditions.

With respect to the second question, we take our results to show that $\overline{A} >$ OFF and $\overline{B} >$ OFF do not entail $\neg(A \wedge B) >$ OFF. As we have explained in Section 1.2, this finding contradicts the predictions of the minimal change requirement as implemented in ordering semantics, no matter what similarity relation among worlds we

[12] In fact, many accounts of exhaustive strengthening assume that if it can occur at all, it can only occur in main clauses; for theoretical issues, see Schlenker (2016).



(a) $\overline{A}$ (b) $\overline{B}$ (c) $\overline{A} \vee \overline{B}$ (d) $A \wedge B$ (e) $\neg[A \wedge B]$

Figure 6: Inquisitive meanings of some simple sentences. ↑↑ represents a world where both switches are up, ↑↓ a world where A is up but B is down, etc. To avoid clutter, only alternatives are depicted.

assume. Hence, we conclude that the interpretation of counterfactuals with complex antecedents challenges the minimal change requirement.

3 Breaking de Morgan's law in conditional antecedents

3.1 Inquisitive semantics

The difference between $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$ finds a natural explanation once we move from a purely truth-conditional notion of meaning to the more fine-grained framework provided by inquisitive semantics (Ciardelli, Groenendijk & Roelofsen 2013). In this framework, the meaning of a sentence φ is given not in terms of truth conditions with respect to possible worlds, but in terms of support conditions with respect to information states, where an information state is modeled as a subset of the set W of possible worlds. The maximal information states supporting a sentence φ are called the alternatives for φ, and the set of alternatives is denoted $\mathrm{Alt}(\varphi)$. A sentence is called inquisitive if it has two or more alternatives, and non-inquisitive if it has only one. The set of worlds where φ is true, denoted $|\varphi|$, is defined as the union of the alternatives for φ. Thus, the inquisitive meaning of a sentence still determines its truth conditions, but the converse is no longer the case: two sentences may very well have the same truth conditions while being associated with different sets of alternatives. This is the case for our counterfactual antecedents $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$. To see why, we need to consider how basic clauses are interpreted in inquisitive semantics, and how disjunction, conjunction, and negation operate in this framework.

First, consider the basic clause *switch A is down*, which we abbreviate as $\overline{A}$. As shown in (9a), this is supported by an information state s in case it follows from the information available in s that switch A is down, that is, in case A is down at each world in s. This in turn means that this clause has a unique alternative, consisting of all those worlds where it is true, as shown in (9b). The same goes for the basic clauses *switch B is down*, *switch A is up*, and *switch B is up*, abbreviated here as $\overline{B}$, A, and B. This is illustrated in Figure 6.

(9) a. $s \models \overline{A}$ iff $s \subseteq \{w \in W \mid \text{switch A is down in } w\}$ iff $s \subseteq |\overline{A}|$
b. $\mathrm{Alt}(\overline{A}) = \{\{w \in W \mid \text{switch A is down in } w\}\} = \{|\overline{A}|\}$

Inquisitive semantics comes with a natural treatment of propositional connectives, obtained by associating these connectives with the natural algebraic operations on the space of inquisitive meanings (see Roelofsen 2013). In particular, disjunction, conjunction, and negation are interpreted by means of the following support clauses:

(10) a. $s \models \varphi \wedge \psi$ iff $s \models \varphi$ and $s \models \psi$
b. $s \models \varphi \vee \psi$ iff $s \models \varphi$ or $s \models \psi$
c. $s \models \neg\varphi$ iff $s \cap t = \varnothing$ for all $t \models \varphi$

We can now verify that in inquisitive semantics, just as in truth-conditional semantics, the sentence *switch A is down* is equivalent with *switch A is not up*, that is, $\overline{A} \equiv \neg A$.[13]

(11) a. $s \models \neg A$ iff $s \cap t = \emptyset$ for all $t \subseteq \{w \in W \mid \text{switch A is up in } w\}$
iff $s \cap \{w \in W \mid \text{switch A is up in } w\} = \varnothing$
iff $s \subseteq \{w \in W \mid \text{switch A is down in } w\}$ iff $s \models \overline{A}$

For our first complex antecedent, *switch A or switch B is down*, analyzed as $\overline{A} \vee \overline{B}$, inquisitive semantics yields two distinct alternatives: the set $|\overline{A}|$ consisting of those worlds where A is down, and the set $|\overline{B}|$ consisting of those worlds where B is down. These alternatives are depicted in Figure 6(c).

(12) a. $s \models \overline{A} \vee \overline{B}$ iff $s \subseteq |\overline{A}|$ or $s \subseteq |\overline{B}|$
b. $\mathrm{Alt}(\overline{A} \vee \overline{B}) = \{|\overline{A}|, |\overline{B}|\}$

Now consider the conjunction *switch A and switch B are both up*, analyzed as $A \wedge B$. Inquisitive semantics yields a unique alternative, consisting of those worlds where both switches are up. This is shown in Figure 6(d).

(13) a. $s \models A \wedge B$ iff $s \subseteq |A|$ and $s \subseteq |B|$
iff $s \subseteq |A \wedge B|$
b. $\mathrm{Alt}(A \wedge B) = \{|A \wedge B|\}$

Finally, for our second complex antecedent, *switch A and switch B are not both up*, analyzed as $\neg(A \wedge B)$, inquisitive semantics yields a unique alternative, consisting of all worlds where the switches are not both up. This is depicted in Figure 6(e).

(14) a. $s \models \neg(A \wedge B)$ iff $s \cap t = \emptyset$ for all $t \subseteq |A \wedge B|$
iff $s \subseteq (W - |A \wedge B|)$
b. $\mathrm{Alt}(\neg(A \wedge B)) = \{W - |A \wedge B|\}$

Since $|\overline{A}| \cup |\overline{B}| = W - |A \wedge B|$, inquisitive semantics predicts that $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$ are true at the same worlds, namely, at those worlds in which one or both switches are down. This is in line with classical logic, and also with the result of Pretest I, as reported in Section 2.3.1. However, these two clauses are assigned different meanings: $\overline{A} \vee \overline{B}$ has two distinct alternatives, whereas $\neg(A \wedge B)$ has only one.

[13] Recall that we are assuming that *up* and *down* are the only possible positions for our switches. The results we obtained in Post-hoc test II justify this assumption.

3.2 Two assumptions for one antecedent

Having explained how $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$ differ in meaning but not in truth conditions, the next step is to explain how this difference ends up affecting the truth conditions of the counterfactuals $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$ in which these clauses are embedded. For this, we adopt an idea due to Alonso-Ovalle (2006, 2009) (see also van Rooij 2006). We assume that a counterfactual antecedent need not always specify a single counterfactual assumption; rather, when an antecedent provides multiple semantic alternatives, as in the case of $\overline{A} \vee \overline{B}$, each of these alternatives constitutes a distinct counterfactual assumption. In order for the whole counterfactual to be true, the consequent must follow on each of these assumptions. Thus, $\overline{A} \vee \overline{B} > \textsc{off}$ is interpreted in effect as the conjunction of $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$, and differently from $\neg(A \wedge B) > \textsc{off}$. This explains the strong similarity between the response pattern of $\overline{A} \vee \overline{B} > \textsc{off}$ and those of $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$.

To implement this idea in our setting, we will apply the general recipe for lifting accounts of counterfactuals into inquisitive semantics described in Ciardelli (2016).[14] The starting point is an arbitrary truth-conditional account of counterfactuals, given in the form of a binary operation $\Rightarrow$ which maps any two propositions p and q to a corresponding conditional proposition $p \Rightarrow q$. Most existing accounts of counterfactuals, including selection function semantics (Stalnaker 1968), ordering semantics (Lewis 1973), premise semantics (Kratzer 1981a), and some causal accounts (Kaufmann 2013, Santorio 2014, forthcoming), can be seen as providing such a map.[15] The lifting recipe interprets a counterfactual sentence, denoted by $\varphi > \psi$, by means of the following support clause.[16]

Definition 1 (Inquisitive lifting of an account of counterfactuals)**.**
$s \models \varphi > \psi$ iff $\forall p \in \mathsf{Alt}(\varphi)\ \exists q \in \mathsf{Alt}(\psi)$ such that $s \subseteq (p \Rightarrow q)$

According to this clause, when φ and ψ are non-inquisitive, that is, $\mathsf{Alt}(\varphi) = \{|\varphi|\}$ and $\mathsf{Alt}(\psi) = \{|\psi|\}$, the conditional $\varphi > \psi$ has a unique alternative, which coincides with the counterfactual proposition delivered by the given base account: $\mathsf{Alt}(\varphi > \psi) = \{|\varphi| \Rightarrow |\psi|\}$.

Except for $\overline{A} \vee \overline{B} > \textsc{off}$, all of the counterfactuals in our experiment have non-inquisitive antecedents and consequents, so they will be interpreted just as they are interpreted by any base account we may choose. As for $\overline{A} \vee \overline{B} > \textsc{off}$, the clause inter-

[14] In Section 6.2 we discuss the reasons why we do not directly adopt Alonso-Ovalle's original account, but turn to the inquisitive lifting recipe instead. In short, that account would not make the right predictions for our data, but for reasons orthogonal to the central idea discussed here.

[15] In each of these accounts, the definition of the conditional proposition $p \Rightarrow q$ makes use of some additional piece of structure: a selection function in Stalnaker (1968), a similarity ordering in Lewis (1973), an ordering source in Kratzer (1981a), and a causal network in Kaufmann (2013). However, our lifting recipe only needs access to the resulting operation on propositions—not to this additional structure.

[16] Throughout the paper, we adopt the convention that the operator $>$ has lower precedence than any other operator. Thus, for instance $A \vee B > C$ should be read as $(A \vee B) > C$, and $A \wedge B > C$ should be read as $(A \wedge B) > C$.

prets it as follows:

$$\begin{aligned} s \models \overline{A} \vee \overline{B} > \textsc{off} \quad & \text{iff} \quad \forall p \in \{|\overline{A}|, |\overline{B}|\}\ \exists q \in \{|\textsc{Off}|\} \text{ such that } s \subseteq (p \Rightarrow q) \\ & \text{iff} \quad s \subseteq |\overline{A}| \Rightarrow |\textsc{Off}| \text{ and } s \subseteq |\overline{B}| \Rightarrow |\textsc{Off}| \\ & \text{iff} \quad s \subseteq (|\overline{A}| \Rightarrow |\textsc{Off}|) \cap (|\overline{B}| \Rightarrow |\textsc{Off}|) \end{aligned}$$

As in the previous cases, the counterfactual as a whole has a unique alternative, namely, the proposition $(|\overline{A}| \Rightarrow |\textsc{Off}|) \cap (|\overline{B}| \Rightarrow |\textsc{Off}|)$. However, this alternative is not the same proposition $|\overline{A} \vee \overline{B}| \Rightarrow |\textsc{Off}|$ that would be delivered by applying the basic truth-conditional account to the sentence. Rather, the basic account is applied twice, once for each disjunct of the antecedent, and the resulting propositions are then intersected. Thus, disjunctive antecedents are interpreted as providing multiple counterfactual assumptions, and $\overline{A} \vee \overline{B} > \textsc{off}$ is predicted to be true just in case both $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ are true.

This means that, at least insofar as truth is concerned, our data will be fully explained if we can find a truth-conditional account of counterfactuals according to which $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ are true, but $\neg(A \wedge B) > \textsc{off}$ and $\neg(A \wedge B) > \textsc{on}$ are not. The inquisitive lifting of this account will still make the same predictions about these cases; moreover, it will predict $\overline{A} \vee \overline{B} > \textsc{off}$ to be true—something that no purely truth-conditional account could do without also making $\neg(A \wedge B) > \textsc{off}$ true.

4 A background theory of counterfactuals

Having explained how $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$ can come apart in their truth values, we now turn to the problem of finding a truth-conditional theory of counterfactuals which predicts that in our scenario, $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ are true, but $\neg(A \wedge B) > \textsc{off}$ and $\neg(A \wedge B) > \textsc{on}$ are not. For this task, one might expect that we can just adopt a standard theory of counterfactuals. Interestingly, however, this is not the case. As we mentioned in Section 1.2, virtually all existing theories of counterfactuals (e.g., Stalnaker 1968, Lewis 1973, Kratzer 1981a, Veltman 2005, Kaufmann 2013) incorporate the minimal change requirement in some form, and this leads them to predict that, if $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ are true, then so is $\neg(A \wedge B) > \textsc{off}$. Therefore, as a result of the minimal change requirement, these theories are not in a position to correctly predict our data, even when disjunctive antecedents are taken care of by the inquisitive lifting recipe.

In this section, we formulate a theory of conditionals which abandons the minimal change requirement, and which fully explains our experimental results when combined with the inquisitive lifting described in Section 3.2. We begin in Section 4.1 by giving an informal description of the theory. In Section 4.2 we introduce some technical notions developed in other causal theories of counterfactuals (Pearl 2000, Schulz 2007, Kaufmann 2013). In Section 4.3 we use these notions to formalize our theory, and we show that the resulting account fully predicts our data.

4.1 The key idea: from minimal change to maximal background

From the perspective of an account that implements the minimal change requirement, what is most surprising about our data is the fact that the counterfactual $\neg(A \wedge B) > \textsc{off}$, repeated below as (15), is not judged true.

(15) If switch A and switch B were not both up, the light would be off.

Let us consider more closely why this is so. When faced with this counterfactual, we appear to reason as follows: if switch A and switch B were not both up, it might be that one of them is down, in which case the light would be off; but it might also be that both of them are down, in which case the light would be on. Hence, no firm conclusion on the state of the light can be reached from our assumption.

If this analysis is correct, then it indicates that in assessing this counterfactual, we consider not just the minimal-change scenarios in which one of the switches is down, but also the non minimal-change scenario in which both switches are down.[17] Intuitively, in this case there is no requirement to minimize departure from actuality: the antecedent invites us to consider situations in which both switches might have different positions, and we feel no pressure at all to limit ourselves to situations which are as similar as possible to the actual one.

To explain this, we propose to dispense with the minimal change requirement, and we replace it by a qualitative distinction between facts that are in the foreground when making a counterfactual assumption and facts that are regarded as background. Background facts are held fixed while making a counterfactual assumption, while foreground facts are allowed to change, and their change is not subject to any minimality requirement.

Crucially, we assume that whether a fact is foregrounded or backgrounded is determined in part by the counterfactual assumption: only facts that are not "called into question" by the assumption can be backgrounded. We assume that a fact f is called into question in case either of the following holds:

1. f contributes to the falsity of a in the actual world;
2. f is dependent on a fact which contributes to the falsity of a.[18]

Given a world and a partition of facts into foreground and background, we say that a counterfactual $\varphi > \psi$, where φ and ψ are non-inquisitive, is true in case ψ follows via causal reasoning from the assumption φ combined with the background facts.

This does not yet allow us to make any specific predictions, since different choices as to what is backgrounded may lead to different truth values. However, we may explain our data by assuming a general preference for maximizing the set of backgrounded facts: that is, we assume that by default, the factual background consists of

[17] In this discussion, we use the terms "similarity" and "minimal change" in a pre-theoretical sense. In the theory that we propose in this section, no corresponding technical notions will be needed.

[18] In this paper, the only type of dependence we consider is causal influence. In Section 5.4, we briefly entertain the possibility of extending this to epistemic dependence.

all and only the facts that are not called into question by the counterfactual assumption.[19]

Given that all facts in the background are held fixed when making a counterfactual assumption, defaulting to a maximal background may be viewed as an analogue of the minimal change requirement. But there is a crucial difference between the two: when it comes to aspects of the world that are not regarded as parts of the background, we assume no requirement to minimize departure from actuality.

Let us see how an account of the kind sketched here provides an explanation for our data. This explanation will then be formalized in Sections 4.2 and 4.3. First consider the counterfactual $\overline{A} > \textsc{off}$. Our counterfactual assumption that A is down directly calls into question the fact that A is up, and indirectly calls into question the fact that the light is on, which is dependent on the fact that A is up. On the other hand, our assumption does not call into question the fact that switch B is up; therefore, this fact will be part of the maximal background for our assumption. Now the assumption that A is down, together with the background fact that B is up, causally implies that the light is off. This explains why the counterfactual $\overline{A} > \textsc{off}$ is mostly judged true. Of course, the situation is completely analogous for the counterfactual $\overline{B} > \textsc{off}$.

Now consider the counterfactuals $\neg(A \wedge B) > \textsc{off}$ and $\neg(A \wedge B) > \textsc{on}$. In this case, the counterfactual assumption that A and B are not both up calls into question both the fact that A is up and the fact that B is up, since these facts are jointly responsible for the falsity of $\neg(A \wedge B)$; the fact that the light is on is called into question as well, since it is dependent on the facts concerning the position of the switches. Since nothing that is called into question by the assumption can be backgrounded, the factual background is empty in this case; thus, no fact about the actual world is retained in making the counterfactual assumption. Now, the assumption $\neg(A \wedge B)$ by itself does not causally imply anything about the state of the light: it does not follow from $\neg(A \wedge B)$ and the way the circuit works that the light is on, and it does not follow that the light is off. This explains why $\neg(A \wedge B) > \textsc{off}$ and $\neg(A \wedge B) > \textsc{on}$ are not judged true in our scenario.

This explanation conveys the fundamental idea of our theory. To turn it into a proper account of our data, we need to make a number of notions formally precise: we need to define what counts as a fact in the actual world, and what it means for a fact to contribute to the falsity of a proposition; we need to specify what it means for a fact to be dependent on another; and we need to say when a conclusion follows via causal reasoning from a set of premises. Our formalization will make use of a number of ideas and notions that have been developed in the literature on causal reasoning (Pearl 2000) and further refined in previous causal approaches to counterfactuals (Schulz 2007, Kaufmann 2013). The next section introduces these notions, which are then employed in 4.3 to formalize our theory.

[19] Importantly, we propose to regard this only as a default choice, and not as an ingredient of the semantics of counterfactuals. In Section 5.1 we suggest, based on some introspective judgments reported to us, that part of the variation in our data might be due to the presence of a second reading of counterfactuals which arises from taking the factual background to be empty. Under this reading, counterfactuals are taken to be general statements about the causal laws, which are wholly independent of the current state of affairs.

4.2 The context for a causal account: causal models

Causal approaches to counterfactuals assume that the evaluation of a counterfactual always takes place in the context of a network of causal relationships that allow for specific causal inferences. Formalizations of this idea within a formal semantic setting have been proposed by Schulz (2007, 2011), Kaufmann (2013) and Santorio (2014, forthcoming). Here we propose our own, which combines elements from these sources.

The core notion is that of a *causal model*, a structure that consists of a set of causal variables, a set of causal laws, and a designated world of evaluation. Formally, a causal model is a triple $M = \langle V, L, w \rangle$ consisting of:

- A set V of *causal variables*, where a causal variable is a partition of the space of possible worlds in at least two blocks. If $X \in V$, a proposition $p \in X$ is called a *setting* of the variable X; if $V' \subseteq V$, a set that contains one setting for each variable in V' is called a *setting of* V'. The value of a variable X at w, denoted X_w, is the unique setting of X that is true at w. Similarly, the value of a set of variables V' at w, denoted V'_w, is the unique setting of V' whose members are all true at w.

  We assume that the variables in V are *independent* from one another, meaning that we require any setting of V to be consistent. Intuitively, this means that the causal variables bear no logical relation to one another, but are only related to one another via the causal laws of the model. For simplicity, we will also assume that the set of causal variables is finite, although this is not essential to our account. In our examples, the causal variables are bipartitions and can therefore be thought of as Boolean variables, but in the general case this need not be so.

- A set L of *laws* encoding causal influence. We represent a law $l \in L$ formally as a tuple $\langle C_l, E_l, m_l \rangle$ where C_l, the *cause set*, is a set of causal variables; E_l, the *effect*, is a causal variable not contained in C_l; and m_l, the *map*, is a partial function from settings of C_l to settings of E_l. Intuitively, the map specifies which causes lead to which effects.

  The *upshot of* l, written $|l|$, is the proposition that is true at those worlds w where m_l is either undefined on the value of C_l at w or maps it to the value of E_l at w. Intuitively, this is the set of worlds at which the causal law in question is obeyed.

- A possible world w, standing for the world of evaluation (typically the actual one). We assume that w respects all causal laws, that is, $w \in |l|$ for all $l \in L$.

In our example, the obvious choice for the set of variables is $V = \{?A, ?B, ?\textsc{On}\}$, where $?A = \{|A|, |\overline{A}|\}$, $?B = \{|B|, |\overline{B}|\}$, and $?\textsc{On} = \{|\textsc{On}|, |\textsc{Off}|\}$. Intuitively, the variables $?A$, $?B$, and $?\textsc{On}$, correspond to the states of the two switches and of the light, respectively. There is only one law; its cause set is $\{?A, ?B\}$, its effect is $?\textsc{On}$, and its map is as follows:

$$\{|A|, |B|\} \mapsto |\textsc{On}| \qquad \{|A|, |\overline{B}|\} \mapsto |\textsc{Off}|$$
$$\{|\overline{A}|, |\overline{B}|\} \mapsto |\textsc{On}| \qquad \{|\overline{A}|, |B|\} \mapsto |\textsc{Off}|$$

The upshot of this law is the proposition $|\textsc{On} \leftrightarrow (A \leftrightarrow B)|$, which is true at a world if the switches have the same position and the light is on, or the switches have different positions and the light is off. Finally, the designated world w is the actual world, in which both switches are up and the light is on.

It is convenient to associate causal models with graphs whose nodes are the causal variables and whose edges indicate relationships of causal influence. Formally, given a causal model $M = \langle V, L, w \rangle$, the *causal graph* of M is the directed graph $G_M = \langle V, E \rangle$ such that E contains an edge from X to Y just in case X is in the cause set of some $l \in L$ whose effect is Y. For example, the causal graph of our example looks as follows:

$$?A \longrightarrow ?\textsc{On} \longleftarrow ?B$$

This represents the fact that the variables $?A$ and $?B$ have a direct causal influence on the variable $?\textsc{On}$, and there are no other causal relations.[20]

4.3 Formalizing our idea: a background theory of counterfactuals

Let us now use the structure provided by a causal model to formulate precisely the account of counterfactuals outlined in Section 4.1. All the definitions in this section assume a causal model $M = \langle V, L, w \rangle$.

The first thing we need to spell out is what counts as a fact in a given state of affairs, and when a fact is dependent on another. We will take a fact to be the value of a causal variable at the given world. In other words, a fact is a true setting of a causal variable.

Definition 2 (Facts)**.**
Recall that, for a variable $X \in V$, we denote by X_w the unique true setting of X at w. The set of facts at w is defined as $\mathcal{F}_w := \{X_w \mid X \in V\}$. We will say that a fact Y_w is dependent on a fact X_w if X is an ancestor of Y in the causal graph of M.[21]

[20]Interesting classes of causal models can be defined by imposing constraints on the associated causal graph. For example, we can restrict to the class of *recursive* causal models, that is, models whose associated graph is acyclic. Clearly, the causal model for our scenario is recursive. However, our general proposal does not require this restriction. Halpern (2013) shows that causal accounts of counterfactuals and ordering semantics come apart on certain nonrecursive models; Santorio (2014, forthcoming) argues that these models can be relevant for natural language and that the empirical predictions of causal accounts surpass those of ordering semantics for these models. Our account inherits these advantages of the causal approach.

[21]Standard theories of conditionals (Stalnaker 1968, Lewis 1973, Kratzer 1981a) make an assumption known as *centering*. This assumption amounts to the fact that any world w is strictly more similar to itself than any other world is. The role of this assumption is to ensure that, in case a is a proposition which is actually true at w, a conditional proposition $a \Rightarrow c$ is true if and only if the consequent c is true. In premise semantics (Kratzer 1981a), this assumption is implemented by requiring that the elements of the ordering source $\mathcal{P}_w$ uniquely characterize the actual world, that is, $\bigcap \mathcal{P}_w = \{w\}$. Similarly, in our setting we may implement the centering assumption by demanding that a world w be uniquely determined by its set of facts, $\bigcap \mathcal{F}_w = \{w\}$. In the setting of our example, this means that our worlds are uniquely determined by the state of the two switches and the state of the light. It is easy to verify that, given the account we are going to spell out, this assumption yields the desired property for counterfactuals with true antecedents.

In our example, in the actual world we have three facts, corresponding to the true settings of our variables: $\mathcal{F}_w = \{|A|, |B|, |\textsc{On}|\}$. The fact $|\textsc{On}|$ is dependent on the facts $|A|$ and $|B|$, and no other causal dependencies hold.

Next, to formulate our ideas we need to specify when a fact contributes to the falsity of an antecedent a. We will say that a fact f contributes to the falsity of a proposition a in case there exists a set $F \subseteq \mathcal{F}_w$ of other facts such that (i) on the basis of F, a might have been true, but (ii) the additional fact f prevents a from being true. More formally, we make the following definition.

Definition 3 (Facts that contribute to the falsity of a proposition)**.**
Let a be a proposition. We say that a fact $f \in \mathcal{F}_w$ contributes to the falsity of a in case there exists some set $F \subseteq \mathcal{F}_w$ such that F is consistent with a, but $F \cup \{f\}$ is not.[22]

Our assumption that the set of causal variables is finite allows us to obtain a useful alternative characterization of this notion. A proof that this characterization is equivalent to the original one is given in Appendix B.

Proposition 1.
A fact $f \in \mathcal{F}_w$ contributes to the falsity of a proposition a in case $f \notin F$ for some maximal set of facts F consistent with a.

Thus, we can check which facts contribute to the falsity of a proposition a by looking at all the maximal sets of facts which are consistent with a. Those facts that belong to all of these sets do not contribute to the falsity of a; the others do.

Definition 4 (Facts responsible for the falsity of a proposition)**.**
We say that a set of facts $F \subseteq \mathcal{F}_w$ is responsible for the falsity of a proposition a in case it contains all and only those facts that contribute to the falsity of a. When F contains just one fact X, we say that X is responsible for the falsity of a.[23]

Let us now consider these notions in our concrete setting. First, consider the proposition that A is down, $|\overline{A}|$. The unique maximal set of facts which is consistent with this proposition is $\{|B|, |\textsc{On}|\}$. Thus, the fact $|A|$ is solely responsible for the falsity of $|\overline{A}|$.

Now consider the proposition that A and B are not both up, $|\neg(A \wedge B)|$. We have two maximal sets of facts that are consistent with this proposition, namely, $\{|A|, |\textsc{On}|\}$ and $\{|B|, |\textsc{On}|\}$. The only fact which belongs to both is $|\textsc{On}|$. Thus, in this case the facts $|A|$ and $|B|$ are jointly responsible for the falsity of $|\neg(A \wedge B)|$.

The next step is to stipulate what facts about the actual state of affairs are called into question when making a counterfactual assumption. We propose that an assumption calls into question those facts that are responsible for its falsity, as well as anything which is dependent on these facts.

Definition 5 (Calling a fact into question)**.**
A proposition a calls into question a fact $f \in \mathcal{F}_w$ if either (i) f contributes to the

[22] We say that F is consistent with a if the intersection of all the propositions in $F \cup \{a\}$ is non-empty.

[23] If the proposition a is actually true at w, the set of facts that are responsible for the falsity of a is empty. Conversely, if we assume centering (see Footnote 21), then whenever a is false at w, the set of facts responsible for the falsity of a is non-empty.

falsity of a, or (ii) f is dependent on some other fact which contributes to the falsity of a.

Consider again our example. The assumption $|\overline{A}|$ calls into question the fact $|A|$, since this fact is responsible for the falsity of $|\overline{A}|$ in the actual world. It also calls into question the fact |On|, which is dependent on $|A|$. On the other hand, it does not call into question the fact $|B|$, since this fact neither contributes to the falsity of $|\overline{A}|$, nor is it dependent on any other fact that does.

Now consider the assumption $|\neg(A \wedge B)|$. This assumption calls into question both $|A|$ and $|B|$, since these two facts contribute to the falsity of $|\neg(A \wedge B)|$. It also calls into question the fact |On|, which is dependent on both $|A|$ and $|B|$. Thus, this assumption calls into question all the facts in our scenario.

We are now going to use the notions introduced to far to put constraints on what facts can be regarded as background for a counterfactual assumption, and thus held fixed in making the assumption and assessing its consequences. We assume that only facts that are not called into question by the assumption can be part of the background.

Definition 6 (Factual background)**.**
A factual background for a proposition a is a set $\mathcal{B}(a)$ of facts which are not called into question by a.

Notice that, for any assumption a, there is a unique maximal factual background for a, namely, the set of all the facts which are not called into question by a. We denote this background by $\mathcal{B}^{max}(a)$.

In our example, the maximal factual background for the assumption that A is down, $|\overline{A}|$, consists of the only fact not called into question by $|\overline{A}|$, the fact that B is up: $\mathcal{B}^{max}(|\overline{A}|) = \{|B|\}$. On the other hand, the maximal factual background for the assumption that A and B are not both up, $|\neg(A \wedge B)|$, is empty, since all facts are called into question by this assumption: $\mathcal{B}^{max}(|\neg(A \wedge B)|) = \emptyset$.

Finally, we need to specify what it means for a conclusion to follow from a set of assumptions by causal reasoning. The idea is to compute the consequences of our assumptions according to the causal laws of our model. However, in general, we need to put constraints on what sets of laws can be held fixed in making a counterfactual assumption a. In particular, we need to get rid of any laws that would contribute to determining the falsity of a. For this, we generalize the notion of intervention in Pearl (2000) to antecedents of arbitrary complexity.

Definition 7 (Law background for a proposition)**.**
Given a proposition a and a law $l \in L$ with effect E, we say that a intervenes on l in case the fact E_w contributes to the falsity of a. The law background for a, denoted $\mathcal{L}(a)$, is the set of all upshots $|l|$ of laws $l \in L$ on which a does not intervene.

In our example, the law background for the assumption $|\overline{A}|$ contains the upshot of the unique law of our circuit: $\mathcal{L}(|\overline{A}|) = \{|\text{On} \leftrightarrow (A \leftrightarrow B)|\}$. This is because the unique fact responsible for the falsity of $|\overline{A}|$ is $|A|$, but no law has the variable $?A$ as its effect. For analogous reasons, we obtain exactly the same law background when we consider the assumptions $|\overline{B}|$ and $|\neg(A \wedge B)|$.

With all these notions in place, we are now ready to specify the truth conditions of a conditional proposition $a \Rightarrow c$ in our account. When we make an assumption a, we first have to decide what facts are to be regarded as background. Given a choice of factual background $\mathcal{B}(a)$, the proposition $a \Rightarrow c$ is true if c is a consequence of a together with the background facts and the background causal laws.

Definition 8 (Truth conditions for counterfactuals)**.**
Let a and c be two propositions, and let $\mathcal{B}(a)$ be a factual background for a. The conditional proposition $a \Rightarrow c$ is true in case $\{a\} \cup \mathcal{B}(a) \cup \mathcal{L}(a)$ logically entails c.[24]

Notice that our account only allows us to make specific predictions for the truth of counterfactuals in combination with a particular strategy for choosing a factual background. We will now show that our majority judgments are accounted for if we assume that the default choice is to make the factual background as large as possible, i.e., to retain all facts that are not directly or indirectly called into question by the counterfactual assumption. Technically, this means choosing for a given assumption a the corresponding maximal factual background, $\mathcal{B}^{max}(a)$.[25]

Definition 9 (Truth conditions under maximal background)**.**
Let a and c be propositions. The conditional proposition $a \Rightarrow c$ is true under a maximal background interpretation in case $\{a\} \cup \mathcal{B}^{max}(a) \cup \mathcal{L}(a)$ logically entails c.[26]

Let us now consider the predictions that this account makes for the counterfactuals in our experiment. First consider the assumption that A was down, $|\overline{A}|$. We saw that the maximal factual background for this assumption is $\mathcal{B}^{max}(|\overline{A}|) = \{|B|\}$, and we saw that the law background contains the upshot of the unique law of our scenario, $|\textsc{On} \leftrightarrow (A \leftrightarrow B)|$. So, to determine whether the counterfactual proposition $|\overline{A}| \Rightarrow |\textsc{Off}|$ is true we need to consider what follows from the following set of propositions:

$$\{\, |\overline{A}|,\ |B|,\ |\textsc{On} \leftrightarrow (A \leftrightarrow B)| \,\}$$

It is easy to see that $|\textsc{Off}|$ follows from this set. Therefore, under a maximal background interpretation, the proposition $|\overline{A}| \Rightarrow |\textsc{Off}|$ is true. Since this proposition is the unique alternative that our inquisitive account assigns to the counterfactual $\overline{A} > \textsc{Off}$, we correctly predict that the counterfactual $\overline{A} > \textsc{off}$ is true.

Of course, the truth of the counterfactual $\overline{B} > \textsc{off}$ is predicted analogously. As for the counterfactual $\overline{A} \vee \overline{B} > \textsc{off}$, we saw in Section 3.2 that it is interpreted by our inquisitive account in Section 3 as equivalent to the conjunction of $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$: thus, this counterfactual is correctly predicted to be true as well.

[24] We say that a set $\mathcal{P}$ of propositions entails a proposition p if p is true in any world where all propositions in $\mathcal{P}$ are true. In other words, $\mathcal{P}$ entails p if the intersection of all propositions in $\mathcal{P}$ is included in p.

[25] This leads to truth conditions that are in line with those in Pearl (2000), although unlike Pearl, we are able to deal with antecedents of arbitrary complexity.

[26] If a is actually true at w, then the maximal background for a is the whole set of facts, $\mathcal{B}^{max}(a) = \mathcal{F}_w$. If we follow standard theories in making the centering assumption we have $\bigcap \mathcal{F}_w = \{w\}$. So, the propositions that follow from a combined with $\mathcal{B}^{max}(a)$ and $\mathcal{L}(a)$ are all and only those propositions that are true at w. So, just as in standard theories, if a is true then $a \Rightarrow c$ has the same truth value as c.

Finally, consider the assumption that switch A and switch B were not both up, $|\neg(A \wedge B)|$. We saw that the maximal factual background for this assumption is empty. We also saw that the law background for this assumption contains the upshot of the unique law of our scenario. So, to determine whether the counterfactual propositions $|\neg(A \wedge B)| \Rightarrow |\textsc{Off}|$ and $|\neg(A \wedge B)| \Rightarrow |\textsc{On}|$ are true, we need to consider what is entailed by the following set of propositions:

$$\{\, |\neg(A \wedge B)|,\ |\textsc{On} \leftrightarrow (A \leftrightarrow B)| \,\}$$

Since neither $|\textsc{Off}|$ nor $|\textsc{On}|$ follows from this set, neither $|\neg(A \wedge B)| \Rightarrow |\textsc{Off}|$ nor $|\neg(A \wedge B)| \Rightarrow |\textsc{On}|$ are predicted to be true. According to our inquisitive account, the first of these propositions is the unique alternative for the counterfactual $\neg(A \wedge B) > \textsc{off}$, and the second is the unique alternative for $\neg(A \wedge B) > \textsc{on}$. Thus, we correctly predict that neither of these counterfactuals is true. Moreover, in this case the prediction does not depend on the assumption of a maximal background interpretation: the empty set is the only law background available for the assumption $|\neg(A \wedge B)|$.[27]

Summing up, then, by combining the background theory of conditionals described in this section with the inquisitive lifting described in Section 3.2 we obtain an account that accurately predicts which of our counterfactuals are true in our scenario, and which ones are not. This is made possible by the combination of (i) the fine-grained view of meaning provided by inquisitive semantics, (ii) an alternative-sensitive account of conditionals, and (iii) a procedure for making counterfactual assumptions which is not constrained by the requirement to minimize departure from the actual world.

5 Extensions

Having accounted for the majority judgments in our experiment, in this section we turn to various additional points that our results raise. We start in Section 5.1 by sketching an account of the minority judgments in terms of a purely causal reading of counterfactuals. We continue in Section 5.2 by pointing out some interesting effects of the order in which the filler sentence and the target sentence were presented, and we suggest a natural explanation of these effects in terms of the factual background parameter in our theory. In Section 5.3, we argue based on introspective evidence that the points we made about counterfactuals can be extended to indicative conditionals. Similarly, in Section 5.4 we argue that these points also concern conditionals whose primary reading is not causal, but epistemic, and we speculate on how our account could be extended to these conditionals.

5.1 Accounting for minority judgments

So far, we have focused on the task of predicting the truth conditions of our sentences in accordance with the judgment of the majority of the experimental participants.

[27] Our scenario is special in that it contains only one law, and our antecedents never intervene on that law. However, our account is also designed to deal with examples in which this is not the case. See Appendix A for such examples drawn from the execution squad scenario in Pearl (2000).

However, our data show that a significant proportion of speakers judged the sentences differently from the majority. Most strikingly, about a third of participants in our main experiment judged the counterfactuals $\overline{A} > \text{OFF}$, $\overline{B} > \text{OFF}$, and $\overline{A} \vee \overline{B} > \text{OFF}$ as indeterminate, rather than true (see Table 3).

While it is possible that some of our data is noise due to careless participants who just happened to judge the filler correctly, not all minority judgments need be interpreted as mistakes or random answers on the part of the subjects. If they were all simple mistakes, we would have to explain why they converge almost unanimously on indeterminate, with hardly any participants judging these sentences false. Rather, we would like to suggest that these judgments may stem from a different—and apparently less salient—reading of our counterfactuals. In particular, based on introspection, it seems plausible that participants who judge $\overline{A} > \text{OFF}$, $\overline{B} > \text{OFF}$, and $\overline{A} \vee \overline{B} > \text{OFF}$ as indeterminate have in mind a purely causal interpretation of counterfactuals. In this interpretation, the current state of the system is disregarded entirely, and only the antecedent and the causal laws are taken into account. In other words, we propose that these participants systematically interpret counterfactuals as general causal statements about the circuit which are not tied to the current situation. As a consequence, they consider all possible positions of the switches that are compatible with the antecedent. The indeterminate judgments then result from the fact that not all of these positions agree on the state of the light.

This explanation is supported by the observation that in Post-hoc test I, where these positions do agree on the state of the light and thus the two readings coincide, the rate of indeterminate judgments dropped to $\sim 10\%$ or less (see Table 4) compared to $\sim 30\%$ in the main experiment.

Our theory captures this purely causal interpretation via the factual background parameter. Whereas the majority interpretation results from maximizing the factual background, the minority interpretation results from minimizing it—i.e., from taking it to be empty (notice that the empty set always counts as a possible background). Under this choice of factual background, our semantics indeed predicts that $a \Rightarrow c$ is true in case c follows from a alone combined with the causal laws on which a does not intervene. In the scenario of our main experiment, there is just one causal law, and none of our antecedents intervene on it. The fact that A is down together with the upshot of this law does not by itself lead to the conclusion that the light is off. Thus, under a purely causal interpretation, $\overline{A} > \text{OFF}$ is not predicted to be true, and similarly for $\overline{B} > \text{OFF}$ and $\overline{A} \vee \overline{B} > \text{OFF}$.

Since our theory only predicts whether a given sentence is true or not, it does not explain on what basis participants who do not judge a sentence true choose between 'indeterminate' and 'false'. Under a purely causal interpretation, lack of a firm conclusion apparently results in an 'indeterminate' rather than 'false' judgment, as witnessed by the responses to $\overline{A} > \text{OFF}$, $\overline{B} > \text{OFF}$, and $\overline{A} \vee \overline{B} > \text{OFF}$ in the main experiment: the 'false' rates for these sentences are dwarfed by the 'indeterminate' rates. By contrast, the 'false' rates for the responses to $\neg(A \wedge B) > \text{OFF}$ and to $\neg(A \wedge B) > \text{ON}$ in the main experiment are substantially higher. In these sentences, the maximal background is empty, so their default and purely causal interpretations coincide. Both lead to a lack of a firm conclusion about the state of the light. Thus, it would appear that a default interpretation that lacks a firm conclusion may result either in a 'false' judg-

Table 7: Order effects in the main experiment: target precedes filler

Sentence	Number	True	(%)	False	(%)	Indet.	(%)
$\overline{A} >$ OFF	125	100	80%	3	2.4%	22	17.6%
$\overline{B} >$ OFF	124	94	75.81%	4	3.22%	26	20.97%
$\overline{A} \vee \overline{B} >$ OFF	185	146	78.92%	9	4.86%	30	16.22%
$\neg(A \wedge B) >$ OFF	193	38	19.69%	82	4.25%	73	37.82%
$\neg(A \wedge B) >$ ON	102	21	20.59%	35	34.31%	46	45.10%

ment or in an 'indeterminate' judgment, while a purely causal interpretation always results in an 'indeterminate' judgment.

It is natural to suppose that 'indeterminate' judgments result from the failure of a homogeneity presupposition to the effect that a counterfactual assumption should lead to a well-determined truth value for the consequent, as proposed by von Fintel (1997). However, the issue of how presupposition failures are reflected in truth value intuitions is a notoriously complex one (on this topic, see von Fintel 2004).

5.2 Order effects

The factual background parameter also allows us to make sense of the observation that in our main experiment, we observed a strong order effect, as shown in Tables 7 and 8. Participants who were presented with the filler sentence $\overline{A} \wedge \overline{B} >$ OFF followed by the target sentence were more likely to judge the target sentence indeterminate than participants who were presented the two sentences in inverse order. This effect was much more pronounced for simple antecedents ($\overline{A} >$ OFF: +27%; $\overline{B} >$ OFF: +23%) and for disjunctive antecedents ($\overline{A} \vee \overline{B} >$ OFF: +22%) than for negated conjunctive antecedents ($\neg(A \wedge B) >$ OFF: +7%; $\neg(A \wedge B) >$ ON: +4%).[28]

Our theory allows us to give a natural explanation of these effects. The fundamental idea of our proposal is that when making a counterfactual assumption, certain facts are foregrounded, i.e., regarded as being at stake, while others are regarded as background and held fixed. To explain the ordering effects, we need only acknowledge that what is regarded as being at stake can be affected by additional factors beyond the given counterfactual assumption. In particular, if a previous sentence invites the

[28]Order effects are highly significant for $\overline{A} >$ OFF ($\chi^2(2, N = 256) = 22.46$, $p < 0.0001$), $\overline{B} >$ OFF ($\chi^2(2, N = 235) = 14.53$, $p = 0.0007$), and $\overline{A} \vee \overline{B} >$ OFF ($\chi^2(2, N = 362) = 21.79$, $p < 0.0001$); borderline significant for $\neg(A \wedge B) >$ OFF ($\chi^2(2, N = 372) = 6.1$, $p = 0.0474$); and not significant for $\neg(A \wedge B) >$ ON ($\chi^2(2, N = 200) = 0.76$, $p = 0.6839$). The main finding of our main experiment is not affected by these order effects, as confirmed in pairwise chi-square tests for data shown in Tables 7 and 8. The patterns are the same in both tables, as indicated by the dashed lines. Comparisons between sentences across blocks within the same table are all highly significant ($p < 0.001$ in both tables), while comparison within blocks are not significant in either table: $\overline{A} >$ OFF vs. $\overline{B} >$ OFF: $\chi^2(2, N = 249) = 0.72$ in Table 7 and $\chi^2(2, N = 242) = 0.97$ in Table 8; $\overline{A} >$ OFF vs. $\overline{A} \vee \overline{B} >$ OFF: $\chi^2(2, N = 310) = 0.53$ in Table 7 and $\chi^2(2, N = 308) = 0.41$ in Table 8; $\overline{B} >$ OFF vs. $\overline{A} \vee \overline{B} >$ OFF: $\chi^2(2, N = 309) = 0.47$ in Table 7 and $\chi^2(2, N = 288) = 0.57$ in Table 8; $\neg(A \wedge B) >$ OFF vs. $\neg(A \wedge B) >$ ON: $\chi^2(2, N = 295) = 0.36$ in Table 7 and $\chi^2(2, N = 277) = 0.84$ in Table 8.

Table 8: Order effects in the main experiment: filler precedes target

Sentence	Number	True	(%)	False	(%)	Indet.	(%)
$\overline{A} >$ OFF	131	69	52.67%	3	2.29%	59	45.04%
$\overline{B} >$ OFF	111	59	53.15%	3	2.70%	49	44.14%
$\overline{A} \vee \overline{B} >$ OFF	177	105	59.32%	5	2.82%	67	37.85%
$\neg(A \wedge B) >$ OFF	179	44	24.58%	54	30.17%	81	45.25%
$\neg(A \wedge B) >$ ON	98	22	22.45%	28	28.57%	48	48.98%

reader to consider a situation in which a certain causal variable is set to a value that is different from its actual one, then the possibility of this variable having a different value may still be salient when the reader considers subsequent sentences. In other words, once a fact has been foregrounded by a sentence, it is more likely to be foregrounded in the interpretation of subsequent sentences.[29]

For instance, suppose a reader is confronted first with the sentence $\overline{A} \wedge \overline{B} >$ OFF, and then with $\overline{A} >$ OFF. The antecedent of $\overline{A} \wedge \overline{B} >$ OFF provides a unique assumption, $|\overline{A} \wedge \overline{B}|$, which calls into question both facts $|A|$ and $|B|$. In other words, in order to interpret $\overline{A} \wedge \overline{B} >$ OFF, one needs to attend to the possibility that the positions of the switches might both be different than they actually are. When interpreting the next sentence, $\overline{A} >$ OFF, some readers may still be attending to this possibility, which leads them to remove $|B|$ from the background, although this fact is not called into question by the assumption $|\overline{A}|$.

This suggests that reading the filler sentence $\overline{A} \wedge \overline{B} >$ OFF first may lead a higher proportion of participants to interpret the sentence $\overline{A} >$ OFF relative to the empty background, leading to a larger proportion of 'indeterminate' judgements. An analogous explanation can be given for the ordering effects that we observed for $\overline{B} >$ OFF and $\overline{A} \vee \overline{B} >$ OFF.

On the other hand, our theory leads us to expect no ordering effect for $\neg(A \wedge B) >$ OFF and $\neg(A \wedge B) >$ ON. This is because the only factual background for the assumption $|\neg(A \wedge B)|$ is the empty set. Therefore, no matter what possibilities previous sentences invite us to consider, this cannot lead to a different choice of factual background for $|\neg(A \wedge B)|$. This explains why the ordering effects for $\neg(A \wedge B) >$ OFF and $\neg(A \wedge B) >$ ON are weak or absent.

5.3 Relevance for indicative conditionals

Since all conditionals in our experiment are counterfactual rather than indicative, our discussion has focused on this specific kind of conditionals. However, we believe that the points made here concern indicative conditionals as well. Indeed, we may modify our target sentences so that instead of talking about what would happen if

[29] This is suggestive of a dynamic view on counterfactuals according to which the set of facts that are foregrounded is continually updated throughout the discourse; see Warmbrōd (1981) and von Fintel (2001) for implementations of related ideas in ordering semantics.

the position of the switches was different, they talk about what will happen if the switches are toggled. Consider the indicative conditionals in (16), which are parallel to our sentences in (1).

(16) a. If we toggle switch A, the light will turn off.
b. If we toggle switch B, the light will turn off.
c. If we toggle switch A or switch B, the light will turn off.
d. If we don't leave switch A and switch B both up, the light will turn off.
e. If we don't leave switch A and switch B both up, the light will remain on.

Based on introspection, we expect the judgments for these sentences to be similar to those we found for the corresponding counterfactuals. If this is correct, the theoretical conclusions we reached about counterfactuals concern conditionals more generally.

5.4 Relevance for epistemic readings of conditionals

In addition to the divide between indicative and counterfactual conditionals, another important distinction is the one between *ontic* and *epistemic* conditionals; this distinction has been emphasized by various scholars, including Lindström & Rabinowicz (1992) and Schulz (2007, 2011). An ontic conditional $A > C$ expresses the fact that intervening on the world so as to make A happen brings C about. An epistemic conditional $A > C$ expresses a disposition, upon learning A, to revise one's beliefs so as to believe that C. As Rott (1999) notes, this is analogous to the distinction between belief update and belief revision known in artificial intelligence (Katsuno & Mendelzon 1991).[30] It is furthermore analogous to the distinction between interventional and observational studies in psychology and medicine (Hagmayer et al. 2007).

While indicative conditionals tend to be interpreted epistemically and counterfactual ones ontically, the correspondence is not perfect. The primary interpretation of the indicative conditionals in (16) is ontic, as these conditionals are naturally taken to be about the effects of a hypothetical change in the world. Conversely, some counterfactuals license an epistemic reading, as attested in a number of questionnaire studies by Sloman & Lagnado (2005) and Rips (2010). Take for example the following minimal pair from Experiment 6 in Sloman & Lagnado (2005):

(17) All rocket ships have two components, A and B. Movement of Component A causes Component B to move. In other words, if A, then B. Both are moving.
a. Suppose Component B were prevented from moving, would Component A still be moving?
b. Suppose Component B were observed to not be moving, would Component A still be moving?

[30] In fact, our observation that the antecedent $\neg(A \wedge B)$ requires us to consider also the possibility that both switches were down is analogous to a problem that arises in the context of belief update (Herzig & Rifi 1999). Disjunctive updates are interpreted inclusively and not exclusively. For example, an object that is dropped on a table whose surface is composed of a black area and a white area may come to rest on the black area, on the white area, or on both. The challenge consists in allowing for the third possibility while guaranteeing that no gratuitous changes occur.

Sloman & Lagnado report that a majority (85%) of participants gave a positive answer to question (17a), while a majority (78%) of participants who were presented the context in (17) gave a negative answer to question (17b). Clearly, the observation mentioned in (17b) that B is not moving does not cause A to stop moving; rather, it allows us to infer that A is not moving; thus, (17b) is an example of a counterfactual conditional whose primary reading is epistemic. By contrast, (17a) is clearly ontic.

While the counterfactuals we focused on in this paper are of the ontic kind, our main observations seem to concern epistemic conditionals as well. This is illustrated by the following example, loosely inspired by a scenario described in Kratzer (1989).

(18) **Context:** The king and the princess like to spend weekends at the castle. The castle has two flags, a green one and a blue one. It is widely known that when the princess is in the castle, one of the flags is raised at random; that when the king is in the castle (either with the princess or by himself), both flags are raised; and that when king and princess are away, both flags are down. On a nearby hill sits the knight, who is in love with the princess and hopes that she is in the castle. He looks over and sees that both flags are up. He sighs as he realizes that he has no way to know if the princess is in the castle, and thinks silently:

(19) If the green flag or the blue flag was down, the princess would be in the castle.

His servant, who is always trying to impress his master, offers:

(20) If the green flag and the blue flag were not both up, the princess would be in the castle.

The knight disagrees:

(21) Were it so easy! If the green flag and the blue flag were not both up, they might both be down. In that case the princess would be away.

It seems to us that the knight is right to believe (19) while rejecting (20). This suggests that de Morgan's law $\neg(\varphi \wedge \psi) \equiv \neg\varphi \vee \neg\psi$ not only fails for ontic counterfactuals, but also for epistemic ones. This poses a novel constraint on theories of belief revision.

The theory of conditionals developed in Section 4, which is based on the notion of causal consequence, is designed to account for the meaning of ontic conditionals. This is the main, if not the only, reading for the sentences in our experiment. Nevertheless, it is conceivable that the core semantics of epistemic and ontic conditionals is more similar than might be expected at first sight given the difference between reasoning about change in the world (e.g. (17a)) and reasoning about belief states (e.g. (17b)). We could interpret epistemic conditionals in the context of a model which is analogous to causal models except that its laws are interpreted as explanatory rather than causal dependencies. Epistemically primitive variables correspond to questions whose answer is immediately accessible, while epistemically dependent variables correspond to questions whose answer is inferred from more basic information according to certain "epistemic laws" which encode generalizations about the world. If this is on the right track, the theory developed here might be applicable to epistemic counterfac-

tuals as well, provided we take the right perspective on what our models represent. Exploring the prospects of such an extension is left as a task for future work.

6 Related work

In this section we relate our work to the existing literature on counterfactuals in two ways: first, we bring our experimental results to bear on issues that have been debated in the literature; second, we compare our positive proposal with other related theories of counterfactuals. We start in Section 6.1 by summarizing the debate on disjunctive antecedents, and we make some observations based on our data. In Section 6.2 we compare our theory to other recent accounts which are similar to ours in that antecedents are not taken to provide a unique proposition as assumption. In Section 6.3 we discuss how our background theory fits within the tradition of premise semantics, and how it departs from it. In Section 6.4 we discuss the issue of inferences from negated conjunctive antecedents, arguing that these have a different origin than inferences from disjunctive antecedents. Finally, in Section 6.5 we strengthen this point by looking at connections between conditionals and modals.

6.1 Simplification of disjunctive antecedents

One of the main points of criticism that have been raised against ordering semantics for counterfactuals (Fine 1975, Nute 1975, Ellis, Jackson & Pargetter 1977) is that it fails to validate the inference pattern known as *simplification of disjunctive antecedents* (SDA).

$$\frac{\varphi \vee \psi > \chi}{\varphi > \chi} \text{ (SDA)}$$

This principle seems intuitively compelling. For instance, Nute (1978) observes that we tend to judge (22a) as false based on the fact that (22b) is false. But this only counts as a reason to reject (22a) if (22b) is indeed a consequence of (22a).

(22) a. If we were to have good weather this summer or if the sun were to grow cold before the end of the summer, we would have a bumper crop.
b. If the sun were to grow cold before the end of the summer, we would have a bumper crop.

Fine (1975), Nute (1975) and Ellis, Jackson & Pargetter (1977) already pointed out that failure to validate SDA is not specific to ordering semantics. Rather, one can prove that no compositional account based on classical logic can validate SDA unless it also validates a general principle of *antecedent strengthening* (AS), which is widely regarded as undesirable in conditional logic.

$$\frac{\varphi > \chi}{\varphi \wedge \psi > \chi} \text{ (AS)}$$

Faced with this problem, various scholars have reacted in different ways. Some (Loewer 1976, McKay & van Inwagen 1977, Lewis 1977) have accepted the equivalence of SDA

with AS, and have taken this as evidence that, in spite of its intuitive appeal, SDA is in fact invalid. This position was supported by some apparent counterexamples to SDA, although the status of such counterexamples has been disputed (see, e.g., Fine 2012b, Willer 2015, Ciardelli 2016).[31]

Other scholars have taken the tension between SDA and AS as a reason to reject compositionality. Notably, Nute (1975) proposed an account in which the semantics of a counterfactual crucially depends on the syntactic form of the antecedent. Loewer (1976) argues that this is too high a price to pay for the validity of SDA. As he points out, "[on] Lewis semantics one has to know how to rank worlds as more or less similar in order to learn the truth conditions of counterfactuals. But in Nute's semantics one has to learn for each formula separately what worlds are relevant in order to evaluate a counterfactual with that formula in the antecedent."

Finally, some scholars have taken this tension as a reason to turn to semantic theories which make more fine-grained distinctions than classical logic affords, and which are capable of disentangling SDA from AS. Our own account falls in this third field, along with Alonso-Ovalle (2009) and Fine (2012b). In Section 6.2, we discuss in some detail the relation between our account and these related proposals.

Here, let us remark that our experimental results provide support for the view that SDA is valid while AS is invalid: on the one hand, the response pattern we found for $\overline{A} \vee \overline{B} > \textsc{off}$ is virtually the same as those of $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$, in line with what SDA would lead us to expect; on the other hand, $\overline{A} > \textsc{off}$ was judged true even though $\overline{A} \wedge \overline{B} > \textsc{off}$, our filler sentence, is clearly false, in contrast with the validity of AS.

6.2 Other accounts of counterfactuals based on fine-grained meanings

One of the main challenges that our account meets is to tease apart the semantics of the two counterfactuals $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$, whose antecedents have the same truth conditions. This is made possible by the combination of the fine-grained notion of meaning provided by inquisitive semantics with a treatment of conditionals which is sensitive to the inquisitive content of the antecedent.

In this respect, our account fits within a family of recent theories of conditionals that assume a fine-grained semantic representation of sentences and make the semantics of conditionals sensitive to more than truth conditions. Proposals in this family include the theory of Alonso-Ovalle (2009), which is based on the framework of alternative semantics; the one of Fine (2012b), which is based on truth-maker semantics; and the one of Willer (2015), which is based on a combination of dynamic semantics and inquisitive semantics.

[31] One such counterexample is due to McKay & van Inwagen (1977):

(i) a. If Spain had fought on either the Axis side or the Allied side, it would have fought on the Axis side.
b. #Therefore, if Spain had fought on the Allied side, it would have fought on the Axis side.

See Footnote 11 in Ciardelli (2016) for a proposal on how to accommodate this kind of counterexample within the inquisitive account of conditionals that we have adopted here.

These accounts use a fine-grained representation of conditional antecedents to validate the principle of simplification of disjunctive antecedents, while blocking full-fledged strengthening of the antecedent. Our experimental results provide further evidence of the need for a fine-grained semantic representation of antecedents: as we argued, a compositional account based only on truth conditions cannot explain the contrast we observed between $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$. However, not all fine-grained accounts are in a position to account for the contrast that we observed. This is because our results are problematic not just for truth-conditional semantics, but for any semantic theory that validates de Morgan's law $\neg(A \wedge B) \equiv \neg A \vee \neg B$. The theories of Fine (2012b) and Willer (2015) do validate this law: therefore, they still lead to the problematic prediction that $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$ are equivalent. Thus, in spite of using a fine-grained semantics, these theories could not account for our findings without a revision of the respective theories of propositional connectives.[32]

Now let us consider the theory of Alonso-Ovalle (2009). This theory is based on a non-standard treatment of disjunction: rather than mapping two propositions p and q to their union $p \cup q$ as in classical logic, disjunction is taken to collect these propositions into a set, delivering $\{p, q\}$. Each element in this set is then treated as a separate counterfactual assumption, and handled by standard ordering semantics.

The fundamental idea of Alonso-Ovalle's theory, that disjunctive antecedents provide multiple assumptions, is also at the basis of our explanation of the contrast between $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$. However, we have implemented this insight in a different way, namely, via the inquisitive lifting recipe developed in Ciardelli (2016). This choice avoids two problems with Alonso-Ovalle's concrete proposal.

First, the semantic framework on which this proposal builds, alternative semantics, has not been equipped with a full-fledged theory of propositional connectives. In fact, Ciardelli, Roelofsen & Theiler (2016) argue that it is difficult, in this framework, to provide a satisfactory treatment of conjunction. Using the inquisitive lifting construction allows us to build on the framework of inquisitive semantics, which comes with a well-developed theory of propositional connectives that shares many of the attractive features of classical logic (see also Roelofsen 2013).

Second, Alonso-Ovalle's theory differs from standard ordering semantics only when disjunction is involved. This means that the argument against ordering semantics that we spelled out in Section 1.2 still applies to this theory: since $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ are true in this scenario, $\neg(A \wedge B) > \textsc{off}$ is predicted to be true, contrary to our experimental findings. Thus, Alonso-Ovalle's theory cannot account for our data, because it builds on ordering semantics, inheriting the problem that we have identified. By contrast, the inquisitive lifting recipe is not tied to a specific base account of counterfactuals, but can be combined with a broad range of accounts. This allowed us to disentangle the problem of dealing with disjunctive antecedents from the problem of determining the right procedure for making counterfactual assumptions. We could therefore address these two problems in turn, and combine the solutions to get a full

[32] In the landscape of truth-maker semantics, a theory which breaks de Morgan's law is the intuitionistic truth-maker semantics of Fine (2014), which is formally related to inquisitive semantics in an interesting way. By assigning different meanings to the antecedents of $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$, this theory provides a suitable starting point for an account of our data. So far, this theory does not seem to have been considered as a starting point for the analysis of counterfactuals, or any other linguistic phenomena.

account of our data.

6.3 Connections with premise semantics

The background theory of counterfactuals that we presented in Section 4 fits within the influential tradition of premise semantics (Kratzer 1981a,b).[33] Kratzer (1981a) lays out the basic idea at the heart of premise semantics as follows:

> 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.

The central notion of ordering semantics is what has come to be known as an *ordering source*, following the terminology in Kratzer (1981b): a contextually determined function g that associates every world w with a set $g(w)$ of propositions, the *premises*. A *premise set* is a subset of $g(w)$. A counterfactual $A > C$ is true just in case for every premise set $P \subseteq g(w)$ that is maximally consistent with A, it is the case that P and A jointly entail C.[34]

Looking at the maximal premise sets among those that are consistent with the antecedent amounts to adding as many facts as consistency permits. This is, in effect, an implementation of the minimal change requirement: in making the counterfactual assumption, we strive to retain as much as possible of the actual state of affairs.

This implementation of the minimal change requirement leads premise semantics to make the same problematic prediction that we discussed in detail for ordering semantics. In fact, Lewis (1981) shows that premise semantics is equivalent to ordering semantics once we allow the similarity relation between worlds to be a weak partial order rather than insisting that it be total. Since the argument we gave in Section 1.2 does not rely on similarity being total, premise semantics still predicts that if both $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ are true, then $\neg(A \wedge B) > \textsc{off}$ is true as well, regardless of the particular ordering source that we consider.

In some respects, our background theory is a version of premise semantics. As in premise semantics, we associate with each world a set of premises, which we call facts or laws; to check whether a counterfactual is true, we consider whether the consequent follows from the antecedent combined with certain premises. Among the existing systems of premise semantics, our system is furthermore similar to the ones of Kaufmann (2013) and Santorio (2014, forthcoming), which like ours use causal models to encode causal connections.

[33] A closely related theory is that of Veltman (1976, 2005). Unlike Kratzer's, it is not formulated in terms of truth conditions; but it is similar to Kratzer's in its workings. In particular, it implements the minimal change requirement in the same way as Kratzer's. The same difference that we will discuss between our background theory and premise semantics also sets our theory apart from Veltman's.

[34] For simplicity, here we focus on the finite case. In the general case, $A > C$ is true if every premise set $P \subseteq g(w)$ that is consistent with A is a subset of some premise set $P' \subseteq g(w)$ that is also consistent with A and such that P' and A jointly entail C. The main difference that we will identify between our theory and standard premise semantics remains in place in the general case.

Nevertheless, there is a fundamental difference between the theory we propose and standard versions of premise semantics. Our analysis departs from the basic idea of the framework, as laid out in the above quote, in that we do not incorporate the minimal change requirement. We propose that there is no general principle forcing us to add to the antecedent "as many facts as consistency permits".[35] Rather, whenever we are faced with a counterfactual assumption, we determine a background of facts which are not called into question, and we hold all these facts fixed. While we do assume a preference for maximizing this background—and thus for avoiding gratuitous changes—we take it to be restricted to those facts that are not called into question by the counterfactual assumption. As we have seen, this allows us to avoid the problematic prediction made by minimal change semantics, and provides an explanation for our experimental findings.

Interestingly, dropping the minimal change requirement also results in a simplification of the account. Whereas in premise semantics we have to consider multiple alternative ways of extending a given counterfactual assumption with a set of premises, in our theory we only have to consider one way of doing so. This is possible because the maximal factual background for a given assumption a is always unique, whereas in general there may not be any single maximal set of premises which is consistent with a.

6.4 Inferences from negated conjunctive antecedents

Any compositional theory of counterfactuals that validates both de Morgan's law $\neg(\varphi \wedge \psi) \equiv \neg\varphi \vee \neg\psi$ and SDA also validates the following principle, which we will refer to as *simplification of negated conjunctive antecedents* (SNCA).

$$\frac{\neg(\varphi \wedge \psi) > \chi}{\neg\varphi > \chi} \text{ (SNCA)}$$

As proponents of such theories point out (Nute 1980, Fine 2012a, Willer 2015), this is a welcome result, since an inference such as (23) does seem sound.

(23) a. If Nixon and Agnew had not both resigned, Ford would never have become President.
b. So if Nixon had not resigned, Ford would never have become President.

Fine (2012a) and Willer (2015) further note that an explanation of SDA as stemming from the presence of the word *or*, such as the one by Alonso-Ovalle (2009), does not account for the validity of this inference. Our theory does not connect the validity of SDA specifically to the presence of the word *or*, but rather to the fact that the

[35] The filtering semantics of Santorio (2014, forthcoming) patterns with our account in this respect: in this account, like in ours, we are not forced to add to the antecedents as many facts as allowed by consistency. Rather, the set of premises is subjected to a "filtering" relative to a given antecedent. While filtering semantics in its existing form still makes wrong predictions about $\neg(A \wedge B) > \text{OFF}$, our background theory can be seen as a proposal to fix this problem by adopting a different filtering recipe. In the absence of any further changes, the modified filtering recipe would then make the wrong predictions about $\overline{A} \vee \overline{B} > \text{OFF}$. Therefore, to obtain a full explanation of our data set, the resulting modification still needs to be combined with a semantic theory that disentangles $\neg(A \wedge B)$ from $\overline{A} \vee \overline{B}$, such as inquisitive semantics.

antecedent is inquisitive.[36] Nevertheless, since negative antecedents are *not* inquisitive, our theory does not explain the inference in (23) in the same way as it explains SDA, namely, as stemming from the fact that the antecedent introduces multiple assumptions. We are thus faced with the challenge of accounting for the validity of the inference in (23) separately.

In this section, we show that our theory does account for this inference. We furthermore argue that the inference patterns licensed by negated conjunctive antecedents are not parallel to those for disjunctive antecedents, contrary to what de Morgan's law would lead us to expect. Therefore, an account that separates the two kinds of inferences, such as ours, is in fact needed.

Let us start from this latter point. The following inference seems to be clearly valid:

(24) a. If Nixon and Agnew had not both resigned, Ford would never have become President. =(23a)
b. So if neither Nixon nor Agnew had resigned, Ford would never have become President.

Schematically, this inference has the following form:

$$\frac{\neg(\varphi \wedge \psi) > \chi}{\neg\varphi \wedge \neg\psi > \chi}$$

Further evidence of the soundness of this inference pattern comes from the fact that, in the context of our experiment, many people object to $\neg(A \wedge B) >$ Off by pointing out that if both switches were down, the light would not be off. Thus, they reject $\neg(A \wedge B) >$ Off based on the fact that $\neg A \wedge \neg B >$ Off is false. But this is only motivated if $\neg(A \wedge B) >$ Off has $\neg A \wedge \neg B >$ Off as a consequence.

Now, if a negated conjunctive antecedent $\neg(\varphi \wedge \psi)$ was treated as equivalent to a disjunctive antecedent $\neg\varphi \vee \neg\psi$, as per de Morgan's law, then we would expect the analogous inference pattern from disjunctive antecedents to be equally acceptable:

$$\frac{\neg\varphi \vee \neg\psi > \chi}{\neg\varphi \wedge \neg\psi > \chi}$$

But this pattern is not valid: in our experiment, (25a) was mostly judged true, while (25b), our filler sentence, is clearly false.

(25) a. If switch A or switch B was down, the light would be off.
b. If switch A and switch B were both down, the light would be off.

Thus, negated conjunctive antecedents license inferences that are unexpected if they are treated on a par with disjunctive antecedents. A compositional account that treats $\neg\varphi \vee \neg\psi$ as equivalent with $\neg(\varphi \wedge \psi)$ must assign the same status to the inference in (24) and to the one in (25), although intuitively only the former is sound.

[36] Another example of conditionals with inquisitive antecedents is given by unconditionals, such as *wherever the party is, I'll go* (Rawlins 2013, Ciardelli 2016). Conditionals whose antecedents contain *any* could also be treated naturally as being inquisitive; see van Rooij (2008) for relevant discussion.

By breaking de Morgan's law, our own account avoids this problem: the inference in (25) is correctly predicted to be invalid on our account, since in the context of our experiment, (25a) is predicted true and (25b) is predicted false. By contrast, we will see that the inference in (24) is predicted to be valid under the natural assumption that Nixon's resignation and Agnew's resignation are treated as facts.

Before coming to this, let us show how the validity of the original inference in (23) is accounted for. We are going to prove the following general fact.

Proposition 2. Let w be a world and let $|A|, |B|$ be two facts at w. Both under a maximal and under a minimal background, if $\neg(A \wedge B) > C$ is true, so is $\neg A > C$.

The key to this result is the following lemma, whose proof is given in Appendix B.

Lemma 1. Suppose $|A|$ and $|B|$ are two facts at w. Then:

- the only fact that is responsible for the falsity of $|\neg A|$ at w is $|A|$;
- the facts that are responsible for the falsity of $|\neg(A \wedge B)|$ at w are $|A|$ and $|B|$;
- the facts that are responsible for the falsity of $|\neg A \wedge \neg B|$ at w are $|A|$ and $|B|$.

It follows immediately from this lemma that anything that is called into question by the assumption $|\neg A|$ is also called into question by the assumption $|\neg(A \wedge B)|$. Since the maximal background for an assumption consists of those facts that are not called into question, it follows that $\mathcal{B}^{max}(|\neg(A \wedge B)|) \subseteq \mathcal{B}^{max}(|\neg A|)$. On the other hand, under a minimal background interpretation, the factual background is taken to be empty for both assumptions. In both cases, we have $\mathcal{B}(|\neg(A \wedge B)|) \subseteq \mathcal{B}(|\neg A|)$.

Now consider the laws. It follows immediately from the lemma that if the assumption $|\neg A|$ intervenes on a law, then $|\neg(A \wedge B)|$ also intervenes on this law. Since the law background for an assumption a consists of the upshots of those laws on which a does not intervene, we have $\mathcal{L}(|\neg(A \wedge B)|) \subseteq \mathcal{L}(|\neg A|)$.

Now suppose that $\neg(A \wedge B) > C$ is true at w. The only alternative for the antecedent is $|\neg(A \wedge B)|$. So, the truth of the counterfactual amounts to the fact that $|C|$ is entailed by the set $\{|\neg(A \wedge B)|\} \cup \mathcal{B}(|\neg(A \wedge B)|) \cup \mathcal{L}(|\neg(A \wedge B)|)$. Since $|\neg A| \subseteq |\neg(A \wedge B)|$, and since $\mathcal{B}(|\neg(A \wedge B)|)$ and $\mathcal{L}(|\neg(A \wedge B)|)$ are included respectively in $\mathcal{B}(|\neg A|)$ and $\mathcal{L}(|\neg A|)$, it follows that $|C|$ is also entailed by $\{|\neg A|\} \cup \mathcal{B}(|\neg A|) \cup \mathcal{L}(|\neg A|)$, which means that $\neg A > C$ is true at w.

This completes the proof of Proposition 2. Thus, provided that the propositions that Nixon resigned and that Agnew resigned are treated as facts in our causal model, the soundness of the inference in (23) is indeed predicted on our account.

The soundness of the inference in (24), which is unexpected for theories that tie (23) to SDA, is predicted on our account in a completely analogous way. That is, the following fact holds.

Proposition 3. Let w be a world and let $|A|, |B|$ be two facts at w. Both under a maximal and under a minimal background, if $\neg(A \wedge B) > C$ is true, so is $(\neg A \wedge \neg B) > C$.

The proof is essentially the same as for Proposition 2. The key ingredient is given by Lemma 1, which states that the propositions $|\neg(A \wedge B)|$ and $|\neg A \wedge \neg B|$ call into question the same facts, namely, $|A|, |B|$, and anything which is dependent on them.

Summing up, then, in this section we have seen that our theory accounts for inferences from negated conjunctive antecedents. In fact, it accounts not just for the inference pattern exemplified by (23), but also for the one exemplified by (24), which is problematic for theories that validate de Morgan's law $\neg(\varphi \wedge \psi) \equiv \neg\varphi \vee \neg\psi$. These inferences do not stem, as SDA does, from the presence of multiple alternatives in the antecedent, but rather from the recipe we described for making counterfactual assumptions. Further evidence for the claim that SDA and SNCA have different origins comes from looking at inferences involving modals, to which we now turn.

6.5 On the relation with free choice in modals

There is wide agreement in the literature that SDA inferences such as (26) are related to free choice inferences under modal operators, illustrated by (27) and (28).

(26) a. If Mr. X wears a top hat or a fedora, he'll blend in with the crowd.
b. So, if he wears a top hat, he'll blend in with the crowd.

(27) a. Mr. X is allowed to wear a top hat or a fedora.
b. So, he is allowed to wear a top hat.

(28) a. Mr. X might be wearing a top hat or a fedora.
b. So, he might be wearing a top hat.

Like SDA, free-choice inferences are not predicted on standard accounts of modal operators; furthermore, like for SDA, making free choice inferences valid leads to unacceptable consequences in these theories, as a result of certain equivalences in classical logic (von Wright 1968, Kamp 1973). The same strategy that we have followed to vindicate SDA has been used to explain the validity of free choice inferences: various scholars have assumed that disjunction introduces multiple propositional alternatives, and that the presence of these alternatives is directly or indirectly responsible for the relevant inferences (Aloni 2003, 2007, Simons 2005, Alonso-Ovalle 2006, Aher & Groenendijk 2015). Thus, free-choice effects provide further motivation for a fine-grained semantic theory such as inquisitive semantics, which is a crucial ingredient of our account.

In addition, the connection between conditionals and modals also allows us to make some interesting observations. In Section 6.4 we argued that inferences from disjunctive antecedents do not have the same source as inferences from negated conjunctive antecedents: the former stem from the presence of semantic alternatives, the latter from the recipe for making counterfactual assumptions.

Looking at free choice effects within the scope of modals provides further evidence for this view. If free choice inferences are linked to the presence of alternatives, we expect them to occur when the prejacent of the modal operator is a disjunction, but not necessarily when it is a negated conjunction. This seems to be precisely the situation: while we saw above that there is a strong parallel between conditionals of the form $(A \vee B) > C$ and modal sentences of the form $\Diamond(A \vee B)$, there is no parallel between conditionals of the form $\neg(A \wedge B) > C$ and modal sentences $\Diamond\neg(A \wedge B)$. The analogues of the inferences in (29a) and (29b) are not valid in the modal setting. Consider (30): one may doubt whether Alice can speak both Dutch and French, yet know for a fact

that she does speak Dutch. Therefore, the inference in (30a) seems unwarranted. In fact, right after (30), the speaker may follow up with an emphatic "but she certainly does speak Dutch!". As for the inference in (30b), it is even more blatantly invalid.[37]

(29) If Bea doesn't speak both Dutch and French, she won't be able to translate.
   a. So, if she doesn't speak Dutch, she won't be able to translate.
   b. So, if she speaks neither Dutch nor French, she won't be able to translate.

(30) Bea might not speak both Dutch and French.
   a. # So, she might not speak Dutch.
   b. # So, she might speak neither Dutch nor French.

If, as it has been assumed, free choice in modals is connected to the presence of multiple alternatives, then on our account it is expected that inferences from disjunctions, but not necessarily from negated conjunctions, carry over to the case of modals.

7 Conclusion

In this paper we reported on a web survey that we conducted to test the truth conditions of certain counterfactual conditionals. The results of this survey indicate that truth-conditionally equivalent antecedents can make different semantic contributions to the interpretation of the conditionals they are part of. Assuming compositionality, this leads to the conclusion that the meaning of these antecedents—and of sentential clauses more generally—should not be identified with their truth conditions. More generally, our data show that de Morgan's law $\neg(\varphi \wedge \psi) \equiv \neg\varphi \vee \neg\psi$ does not hold in natural language: a compositional account of our data requires a theory of propositional connectives that assigns different semantic values to $\neg(A \wedge B)$ and $\neg A \vee \neg B$.

We have shown that a natural explanation of our data is available in inquisitive semantics: the inquisitive account of propositional connectives distinguishes $\neg(A \wedge B)$ from $\neg A \vee \neg B$, by associating the first clause with a single semantic alternative, and the second clause with two distinct alternatives. The inquisitive lifting recipe of Ciardelli (2016), which treats each alternative for the antecedent as a separate counterfactual assumption, then explains how this difference affects the truth conditions of the conditionals in which these two clauses are embedded.

Our findings also challenge the widespread view that making a counterfactual assumption requires minimizing the amount of change with respect to the actual state of affairs. We have seen that, no matter what exactly is taken to count as a minimal change in our scenario, our data cannot be accounted for. Another way to put it is this: on a theory that implements the minimal change requirement, the majority judgments that we found are predicted to be logically inconsistent.

We have proposed that in making a counterfactual assumption, there is no general requirement to minimize changes; rather, certain facts are regarded as background for the assumption, and held fixed in the counterfactual scenario. We have furthermore

[37] Notice that *might* always takes scope above ordinary negation: the sentence *Alice might not be home* can only mean that it is possible that Alice is not home. This ensures that the logical form of the sentences in (30) is indeed $\Diamond\neg(D \wedge F)$, $\Diamond\neg D$, and $\Diamond(\neg D \wedge \neg F)$.

assumed that a fact is by default viewed as background unless it is called into question by the counterfactual assumption. We have developed a formal account based on this view, and we have shown that this account, when suitably combined with inquisitive semantics, predicts our majority judgments.

Appendix

A Illustration of our theory with intervention on causal laws

In this appendix we illustrate the workings of the theory we developed in Section 4 by looking at the interpretation of two counterfactuals in a scenario taken from Pearl (2000). The scenario involves two riflemen who are preparing to shoot a prisoner upon receiving a signal from the squad's officer, who is waiting for the court order. Let $V = \{?C, ?O, ?A, ?B, ?D\}$ where $?C = \{C, \overline{C}\}$ symbolizes whether the court orders the execution, $?O = \{O, \overline{O}\}$ whether the officer transmits the order to the riflemen, $?A = \{A, \overline{A}\}$ and $?B = \{B, \overline{B}\}$ whether the riflemen shoot, and $?D = \{D, \overline{D}\}$ whether the prisoner dies. Let the causal graph be such that D depends on A and B, which each depend on O, which depends on C, which does not depend on anything. Let L be laws whose upshots are $C \leftrightarrow O$, $O \leftrightarrow A$, $O \leftrightarrow B$, and $(A \vee B) \leftrightarrow D$. Suppose in the actual world, the court gives the order, the officer transmits it to both riflemen, they both shoot, and the prisoner dies; this means that $\mathcal{F}_w = \{C, O, A, B, D\}$.

Example 1. Consider the counterfactual "if rifleman A hadn't shot, the prisoner would still have died", represented as $\neg A > C$.[38] The only fact that contributes to the falsity of $\neg A$ is A; only D is dependent on it; so $\neg A$ calls into question A and D; the remaining facts, B, C, and O, form the maximal factual background of $\neg A$. The law background for $\neg A$ consists of the upshots of all the laws except $O \leftrightarrow A$. Now, since the maximal factual background contains B, and since the law background contains the law with upshot $(A \vee B) \leftrightarrow D$, we have that $\{\neg A\} \cup \mathcal{B}^{max}(\neg A) \cup \mathcal{L}(\neg A)$ entails D. Thus, on the maximal background interpretation, the counterfactual is indeed true.

Example 2. In the same scenario, consider the counterfactual "if riflemen A and B had not both shot, the prisoner would still have died", represented as $\neg(A \wedge B) > D$. We expect this to be judged false or indeterminate, since if the riflemen had not both shot they might have both refrained from shooting. The only facts that contribute to the falsity of $\neg(A \wedge B)$ are A and B; only D is dependent on either of these facts; so, $\neg(A \wedge B)$ calls into question A, B, and D; the remaining facts, C and O, form the maximal factual background of $\neg(A \wedge B)$. The law background for $\neg(A \wedge B)$ consists of the upshots of all the laws except $O \leftrightarrow A$ and $O \leftrightarrow B$. Since both these laws are excluded, there is no way to conclude from the remaining information that the prisoner would have died. Thus, the counterfactual is not predicted to be true.

[38] In these examples, we blur the distinction between a clause A and the proposition $|A|$, for the sake of readability. In general, however, this distinction is important in our account: since a single antecedent may give rise to multiple counterfactual assumptions, evaluating a counterfactual $A > C$ does not always amount to computing the proposition $|A| \Rightarrow |C|$. In particular, $|(A \vee B) > C|$ is in general distinct from $|A \vee B| \Rightarrow |C|$.

B Proofs of mathematical results

Proof of Proposition 1. Suppose f contributes to the falsity of a at w. This means that there is some $F \subseteq \mathcal{F}_w$ such that F is consistent with a but $F \cup \{f\}$ is not. Since the set V of causal variables is finite, the set $\mathcal{F}_w$ of facts is finite too. Therefore, F can be extended to some set $F' \subseteq \mathcal{F}_w$ which is maximal among the subsets of $\mathcal{F}_w$ consistent with a. Since $F \subseteq F'$ and $F \cup \{f\}$ is inconsistent with a, *a fortiori* the set $F' \cup \{f\}$ is inconsistent with a. Since F' *is* consistent with a, we must have $F' \cup \{f\} \neq F'$, which implies $f \notin F'$. So, for some maximal set of facts F' consistent with a, $f \notin F'$.

Conversely, suppose f is does not contribute to the falsity of a at w. Now take a set of facts $F \in \mathcal{F}_w$ which is maximal among whose consistent with a. Since f does not contribute to the falsity of a, and since and F is consistent with a, we have that $F \cup \{f\}$ must be consistent with a as well. Since F is maximal among the sets of facts consistent with a, we cannot have $F \cup \{f\} \supset F$: we must then have $F \cup \{f\} = F$, which means that $f \in F$. Since F was an arbitrary set of facts which is maximally consistent with a, this shows that s is included in all such sets. □

Proof of Lemma 1. Suppose that $|A|$ and $|B|$ are facts in our model. Since inquisitive semantics coincides with classical logic as far as the truth-conditions of the propositional connectives are concerned, we have $|\neg A| = \overline{|A|}$, $|\neg(A \wedge B)| = \overline{|A| \cap |B|}$, and $|\neg A \wedge \neg B| = \overline{|A|} \cap \overline{|B|}$. Thus, our lemma will be established if we can prove the following three claims for any facts f and g at a world w:

1. the only fact that is responsible for the falsity of $\overline{f}$ at w is f;
2. the facts that are responsible for the falsity of $\overline{f \cap g}$ at w are f and g;
3. the facts that are responsible for the falsity of $\overline{f} \cap \overline{g}$ at w are f and g.

Let us establish these claims in turn.

1. Consider the set of facts $F := \mathcal{F}_w - \{f\}$. We claim that this is the only maximal set of facts consistent with $\overline{f}$. Let us prove this.
   - F is consistent with $\overline{f}$. Let X be the causal variable such that $f = X_w$, and let f' be a different setting of X. Then $F \cup \{f'\}$ is a setting of V, and so it is consistent, by our assumptions that the variables in V are logically independent from one another.[39] Now, since the settings for a variable form a partition, we have $f' \subseteq \overline{f}$. Thus, $F \cup \{\overline{f}\}$ is consistent as well, which means that F is consistent with $\overline{f}$.
   - Clearly, F is maximal among the sets consistent with $\overline{f}$: the only proper superset of F is is $\mathcal{F}_w$, which contains f and is therefore inconsistent with $\overline{f}$.

[39] Recall from Section 4.2 that we assume that the causal variables in V are logically independent from one another. Technically, what this means is that any setting of V is logically consistent.

- F is the unique maximal set of facts consistent with $\overline{f}$. To see this, suppose H is a set of facts consistent with $\overline{f}$: then f cannot belong to H, so $H \subseteq F$.

By Proposition 1, the facts that are responsible for the falsity of $\overline{f}$ are all and only those that are not included in F. By definition, f is the only such fact.

2. Consider the set of facts $F := \mathcal{F}_w - \{f\}$ and $G := \mathcal{F}_w - \{g\}$. We claim that these are the unique maximal sets of facts consistent with $\overline{f \cap g}$. Let us show this.

   - F and G are consistent with $\overline{f \cap g}$. First consider F. Suppose $f = X_w$, and let f' be a different setting of X. Then, $F \cup \{f'\}$ is a setting of V, and so it is consistent by the independence of V. Since the settings for a variable form a partition, we have $f' \subseteq \overline{f} \subseteq \overline{f \cap g}$. Thus, $F \cup \{\overline{f \cap g}\}$ is consistent as well, which means that F is consistent with $\overline{f \cap g}$. The argument is similar for G.
   - F and G are maximal among the set of facts consistent with $\overline{f \cap g}$. This is obvious, since the only proper superset of either F or G is the full set $\mathcal{F}_w$, which contains both f and g and is therefore not consistent with $\overline{f \cap g}$.
   - F and G are the unique maximal set of facts consistent with $\overline{f \cap g}$. To prove this, it suffices to show that any set of facts H which is consistent with $\overline{f \cap g}$ is included either in F or in G. So, suppose H is consistent with $\overline{f \cap g}$. Then H cannot include both f and g: if H does not include f, then $H \subseteq F$, while if H does not include g, $H \subseteq G$.

   By Proposition 1, the facts that contribute to the falsity of $\overline{f \cap g}$ are those that are not included in F, or not included in G. Clearly, the only such facts are f and g.

3. Consider the set of facts $H := \mathcal{F} - \{f, g\}$. We claim that H is the unique maximal set of fact consistent with $\overline{f} \cap \overline{g}$. Let us show this.

   - H is consistent with $\overline{f} \cap \overline{g}$. To see this, suppose $f = X_w$ and $g = Y_w$. Let f' and g' be different settings of the variables f and g. Then the set $H \cup \{f', g'\}$ is a setting of V, and thus it is consistent by the independence of the set V of causal variables. Since the settings of a causal variable form a partition, we have $f' \subseteq \overline{f}$ and $g' \subseteq \overline{g}$. Thus, the set $H \cup \{\overline{f}, \overline{g}\}$ is consistent as well. But this is equivalent to $H \cup \{\overline{f} \cap \overline{g}\}$ being consistent, which by definition amounts to H being consistent with $\overline{f} \cap \overline{g}$.
   - H is maximal among the set of facts consistent with $\overline{f} \cap \overline{g}$, since any proper superset of H must contain either f or g, and must therefore be inconsistent with $\overline{f} \cap \overline{g}$.
   - H is the unique maximal set of facts consistent with $\overline{f} \cap \overline{g}$. To see this, consider an arbitrary set of facts H' which is consistent with $\overline{f} \cap \overline{g}$. Then, H' cannot contain either f or g, which means that $H' \subseteq H$.

By Proposition 1, the facts responsible for the falsity of $\overline{f} \cap \overline{g}$ are those that are not included in H, namely, f and g. □

Acknowledgments

Champollion, Ciardelli & Zhang (2016) is an earlier and shorter version of this paper. For comments and discussion, we thank Luis Alonso-Ovalle, Rebekah Baglini, Justin Bledin, Joseph DeVeaugh-Geiss, Kit Fine, Johannes Marti, Robert van Rooij, Paolo Santorio, Katrin Schulz, Anna Szabolcsi, Frank Veltman, Malte Willer, and audiences at SALT 26, at the Fourth Workshop on Natural Language and Computer Science (NLCS 2016), and in Utrecht, Paris, and Göttingen. Special thanks to Floris Roelofsen. Ivano Ciardelli gratefully acknowledges financial support form the Netherlands Organization for Scientific research (NWO). Lucas Champollion gratefully acknowledges financial support from the University Research Challenge Fund (URCF) at New York University.

References

Aher, Martin & Jeroen Groenendijk. 2015. Deontic and epistemic modals in suppositional [inquisitive] semantics. In Eva Csipak & Hedde Zeijlstra (eds.), *Sinn und Bedeutung 19*, 2–19. Göttingen, Germany. https://www.uni-goettingen.de/en/proceedings/521400.html.

Aloni, Maria. 2003. Free choice in modal contexts. In Matthias Weisgerber (ed.), *Sinn und Bedeutung 7*, 25–37. Konstanz, Germany: Fachbereich Sprachwissenschaft, Universität Konstanz. http://ling.uni-konstanz.de/pages/conferences/sub7/proceedings/download/sub7_aloni.pdf.

Aloni, Maria. 2007. Free choice, modals, and imperatives. *Natural Language Semantics* 15(1). 65–94. http://dx.doi.org/10.1007/s11050-007-9010-2.

Aloni, Maria. 2016. Disjunction. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, Summer 2016. http://plato.stanford.edu/archives/sum2016/entries/disjunction/.

Alonso-Ovalle, Luis. 2006. *Disjunction in alternative semantics*. Amherst, MA: University of Massachusetts Amherst PhD thesis. http://scholarworks.umass.edu/dissertations/AAI3242324/.

Alonso-Ovalle, Luis. 2009. Counterfactuals, correlatives, and disjunction. *Linguistics and Philosophy* 32(2). 207–244. http://dx.doi.org/10.1007/s10988-009-9059-0.

Champollion, Lucas, Ivano Ciardelli & Linmin Zhang. 2016. Breaking de Morgan's law in counterfactual antecedents. *26th Semantics and Linguistic Theory Conference (SALT 26)* 26. 304–324. http://dx.doi.org/10.3765/salt.v26io.3800.

Ciardelli, Ivano. 2016. Lifting conditionals to inquisitive semantics. *26th Semantics and Linguistic Theory Conference (SALT 26)*. 732–752. http://dx.doi.org/10.3765/salt.v26io.3811.

Ciardelli, Ivano, Jeroen Groenendijk & Floris Roelofsen. 2013. Inquisitive semantics: A new notion of meaning. *Language and Linguistics Compass* 7(9). 459–476. http://dx.doi.org/10.1111/lnc3.12037.

Ciardelli, Ivano, Floris Roelofsen & Nadine Theiler. 2016. Composing alternatives. *Linguistics and Philosophy*. 1–36. http://dx.doi.org/10.1007/s10988-016-9195-2.

Ellis, Brian, Frank Jackson & Robert Pargetter. 1977. An objection to possible-world semantics for counterfactual logics. *Journal of Philosophical Logic* 6(1). 355–357. http://dx.doi.org/10.1007/bf00262069.

Erlewine, Michael Yoshitaka & Hadas Kotek. 2016. A streamlined approach to online linguistic surveys. *Natural Language and Linguistic Theory* 34(2). 481–495. http://dx.doi.org/10.1007/s11049-015-9305-9.

Fine, Kit. 1975. Critical notice. *Mind* 84(335). 451–458. http://dx.doi.org/10.1093/mind/LXXXIV.1.451.

Fine, Kit. 2012a. A difficulty for the possible worlds analysis of counterfactuals. *Synthese* 189(1). 29–57. http://dx.doi.org/10.1007/s11229-012-0094-y.

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

Fine, Kit. 2014. Truth-maker semantics for intuitionistic logic. *Journal of Philosophical Logic* 43(2-3). 549–577. http://dx.doi.org/10.1007/s10992-013-9281-7.

von Fintel, Kai. 1997. Bare plurals, bare conditionals, and *only*. *Journal of Semantics* 14(1). 1–56. http://dx.doi.org/10.1093/jos/14.1.1.

von Fintel, Kai. 2001. Counterfactuals in a dynamic context. In Michael Kenstowicz (ed.), *Ken Hale: a life in language*, vol. 36, 123–152. Cambridge, MA: MIT.

von Fintel, Kai. 2004. Would you believe it? the king of France is back! Presuppositions and truth-value intuitions. In Anne Bezuidenhout & Marga Reimer (eds.), *Descriptions and beyond: an interdisciplinary collection of essays on definite and indefinite descriptions and other related phenomena*, 315–341. Oxford, UK: Oxford University Press.

Fox, Danny. 2007. Free choice and the theory of scalar implicatures. In Uli Sauerland & Penka Stateva (eds.), *Presupposition and implicature in compositional semantics*, 71–120. London, UK: Palgrave Macmillan. http://dx.doi.org/10.1057/9780230210752.

Groenendijk, Jeroen & Martin Stokhof. 1990. Dynamic Montague grammar. In László Kálmán & László Pólos (eds.), *Papers from the 2nd symposium on logic and language*, 3–48. Budapest, Hungary: Akadémiai Kiadó.

Hagmayer, York, Steven Sloman, David Lagnado & Michael R. Waldmann. 2007. Causal reasoning through intervention. In Alison Gopnik & Laura Schulz (eds.), *Causal learning: psychology, philosophy, and computation*, 86–100. Oxford,UK: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195176803.003.0007.

Halpern, Joseph Y. 2013. From causal models to counterfactual structures. *The Review of Symbolic Logic* 6(2). 305–322. http://dx.doi.org/10.1017/s1755020312000305.

Hamblin, Charles J. 1973. Questions in Montague English. *Foundations of Language* 10(1). 41–53. http://www.jstor.org/stable/25000703.

Heim, Irene. 1982. *The semantics of definite and indefinite noun phrases*. Amherst, MA: University of Massachusetts PhD thesis. http://semanticsarchive.net/Archive/jA2YTJmN.

Heim, Irene & Angelika Kratzer. 1998. *Semantics in Generative Grammar*. Oxford, UK: Blackwell Publishing.

Herzig, Andreas & Omar Rifi. 1999. Propositional belief base update and minimal change. *Artificial Intelligence* 115(1). 107–138. http://dx.doi.org/10.1016/s0004-3702(99)00072-7.

Horn, Laurence. 1972. *On the semantic properties of logical operators in English*. Los Angeles, CA: University of California PhD thesis.

Kamp, Hans. 1973. Free choice permission. *Proceedings of the Aristotelian Society* 74(1). 57–74. http://dx.doi.org/10.1093/aristotelian/74.1.57.

Kamp, Hans. 1981. A theory of truth and semantic representation. In Jeroen Groenendijk, Theo Janssen & Martin Stokhof (eds.), *Formal methods in the study of language*, vol. 135 (Mathematical Center Tracts), 277–322. Amsterdam, Netherlands. http://dx.doi.org/10.1002/9780470758335.ch8.

Katsuno, Hirofumi & Alberto O. Mendelzon. 1991. On the difference between updating a knowledge base and revising it. In James Allen, Richard E. Fikes & Erik Sandewall (eds.), *Principles of knowledge representation and reasoning: proceedings of the second international conference*, 387–394. San Mateo, CA: Morgan Kaufmann.

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

Kratzer, Angelika. 1981a. Partition and revision: the semantics of counterfactuals. *Journal of Philosophical Logic* 10(2). 201–216. http://dx.doi.org/10.1007/bf00248849.

Kratzer, Angelika. 1981b. The notional category of modality. In Hans-Jürgen Eikmeyer & Hannes Rieser (eds.), *Words, worlds, and contexts: new approaches in word semantics*, vol. 6 (Research in text theory), 38–74. Berlin, Germany: de Gruyter. http://dx.doi.org/10.1002/9780470758335.ch12.

Kratzer, Angelika. 1989. An investigation of the lumps of thought. *Linguistics and Philosophy* 12(5). 607–653. http://dx.doi.org/10.1007/bf00627775.

Kratzer, Angelika & Junko Shimoyama. 2002. Indeterminate pronouns: The view from Japanese. In Yukio Otsu (ed.), *3rd Tokyo conference on psycholinguistics*, 1–25.

Lewis, David. 1973. *Counterfactuals*. Oxford, UK: Blackwell.

Lewis, David. 1977. Possible-world semantics for counterfactual logics: a rejoinder. *Journal of Philosophical Logic* 6(1). 359–363. http://dx.doi.org/10.1007/bf00262070.

Lewis, David. 1979. Counterfactual dependence and time's arrow. *Noûs* 13(4). 455–476. http://dx.doi.org/10.2307/2215339.

Lewis, David. 1981. Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic* 10(2). 217–234. http://dx.doi.org/10.1007/bf00248850.

Lifschitz, Vladimir. 1990. Frames in the space of situations. *Artificial Intelligence* 46(3). 365–376. http://dx.doi.org/10.1016/0004-3702(90)90021-q.

Lindström, Sten & Wlodzimierz Rabinowicz. 1992. The Ramsey test revisited. *Theoria* 58(2-3). 131–182. http://dx.doi.org/10.1111/j.1755-2567.1992.tb01138.x.

Loewer, Barry. 1976. Counterfactuals with disjunctive antecedents. *The Journal of Philosophy* 73(16). 531–537. http://dx.doi.org/10.2307/2025717.

McKay, Thomas & Peter van Inwagen. 1977. Counterfactuals with disjunctive antecedents. *Philosophical Studies* 31(5). 353–356. http://dx.doi.org/10.1007/bf01873862.

Menzies, Peter. 2014. Counterfactual theories of causation. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, Spring 2014. http://plato.stanford.edu/archives/spr2014/entries/causation-counterfactual/.

Nute, Donald. 1975. Counterfactuals and the similarity of words. *The Journal of Philosophy* 72(21). 773–778. http://dx.doi.org/10.2307/2025340.

Nute, Donald. 1978. Simplification and substitution of counterfactual antecedents. *Philosophia* 7(2). 317–325. http://dx.doi.org/10.1007/bf02378818.

Nute, Donald. 1980. Conversational scorekeeping and conditionals. *Journal of Philosophical Logic* 9(2). 153–166. http://dx.doi.org/10.1007/bf00247746.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/cbo9780511803161.

Rawlins, Kyle. 2013. (Un)conditionals. *Natural Language Semantics* 21(2). 111–178. http://dx.doi.org/10.1007/s11050-012-9087-0.

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

Roelofsen, Floris. 2013. Algebraic foundations for the semantic treatment of inquisitive content. *Synthese* 190(1). 79–102. http://dx.doi.org/10.1007/s11229-013-0282-4.

van Rooij, Robert. 2006. Free choice counterfactual donkeys. *Journal of Semantics* 23(4). 383–402. http://dx.doi.org/10.1093/jos/ffl004.

van Rooij, Robert. 2008. Towards a uniform analysis of *any*. *Natural Language Semantics* 16(4). 297–315. http://dx.doi.org/10.1007/s11050-008-9035-1.

Rooth, Mats. 1985. *Association with focus*. Amherst, MA: University of Massachusetts Amherst PhD thesis. http://scholarworks.umass.edu/dissertations/AAI8509599/.

Rooth, Mats. 1996. Focus. In Shalom Lappin (ed.), *Handbook of contemporary semantic theory*, 271–297. Oxford, UK: Blackwell Publishing. http://dx.doi.org/10.1111/b.9780631207498.1997.00013.x.

Rott, Hans. 1999. Moody conditionals: Hamburgers, switches, and the tragic death of an American president. In Jelle Gerbrandy, Maarten Marx, Maarten de Rijke & Yde Venema (eds.), *JFAK: Essays dedicated to Johan van Benthem on the occasion of his 50th birthday*, 98–112. Amsterdam, Netherlands: Amsterdam University Press.

Santorio, Paolo. 2014. Filtering semantics for counterfactuals: bridging causal models and premise semantics. *24th Semantics and Linguistic Theory Conference (SALT 24)*. 494–513. http://dx.doi.org/10.3765/salt.v24i0.2430.

Santorio, Paolo. Forthcoming. Interventions in premise semantics. *Philosophers' Imprint*.

Schlenker, Philippe. 2016. The semantics–pragmatics interface. In Maria Aloni & Paul Dekker (eds.), *Cambridge handbook of formal semantics* (Cambridge Handbooks in Language and Linguistics), chap. 22, 664–727. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/cbo9781139236157.023.

Schulz, Katrin. 2007. *Minimal models in semantics and pragmatics: Free choice, exhaustivity, and conditionals*. Amsterdam, Netherlands: University of Amsterdam PhD thesis. http://hdl.handle.net/11245/1.272471.

Schulz, Katrin. 2011. ''If you'd wiggled A, then B would've changed". *Synthese* 179(2). 239–251. http://dx.doi.org/10.1007/s11229-010-9780-9.

Simons, Mandy. 2005. Dividing things up: the semantics of or and the modal/or interaction. *Natural Language Semantics* 13(3). 271–316. http://dx.doi.org/10.1007/s11050-004-2900-7.

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

Spector, Benjamin. 2007. Aspects of the pragmatics of plural morphology: On higher-order implicatures. In Uli Sauerland & Penka Stateva (eds.), *Presuppositions and implicature in compositional semantics*, 243–281. London, UK: Palgrave. http://dx.doi.org/10.1057/9780230210752.

Stalnaker, Robert C. 1968. A theory of conditionals. In Nicholas Rescher (ed.), *Studies in logical theory*, 98–113. Oxford, UK: Blackwell. http://dx.doi.org/10.1007/978-94-009-9117-0_2.

Veltman, Frank. 1976. Prejudices, presuppositions, and the theory of counterfactuals. In *Amsterdam papers in formal grammar. first Amsterdam Colloquium*, 248–282. University of Amsterdam. http://hdl.handle.net/11245/1.428635.

Veltman, Frank. 2005. Making counterfactual assumptions. *Journal of Semantics* 22(2). 159–180. http://dx.doi.org/10.1093/jos/ffh022.

Warmbrōd, Ken. 1981. Counterfactuals and substitution of equivalent antecedents. *Journal of Philosophical Logic* 10(2). 267–289. http://dx.doi.org/10.1007/bf00248853.

Willer, Malte. 2015. Simplifying counterfactuals. In Thomas Brochhagen, Floris Roelofsen & Nadine Theiler (eds.), *20th Amsterdam Colloquium*, 428–437. ILLC Publications.

von Wright, Georg Henrik. 1968. *An essay in deontic logic and the general theory of action*. Vol. 21 (Acta Philosophica Fennica). Amsterdam, Netherlands: North-Holland.