Two switches in the theory of counterfactuals

A study of truth conditionality and minimal change

Ivano Ciardelli
ivano.ciardelli@lmu.de

Linmin Zhang
linmin.zhang@nyu.edu

Lucas Champollion
champollion@nyu.edu

This draft compiled on December 9, 2017

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. 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, experimental semantics, crowdsourcing survey, disjunctive antecedents, inquisitive semantics, minimal change semantics, ordering semantics, 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 wiring diagram in Figure 1 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…

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

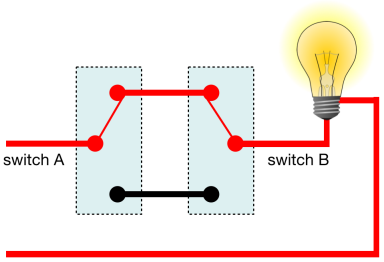


Figure 1: A multiway switch

(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, the nature of sentence meaning, and the interpretation of counterfactuals. 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 refer to the sentences in (1) by the mnemonic labels that appear next to them. These labels are based on the following conventions: A and $\overline{A}$ stand respectively for '*switch A is up*' and '*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. We take $>$ to have lower precedence than other operators; thus, for example, $\overline{A} \vee \overline{B} > \textsc{off}$ should be read as $(\overline{A} \vee \overline{B}) > \textsc{Off}$. 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 semantic 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. In such frameworks, 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 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 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 $\neg(A \wedge B)$ is 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, $\overline{A} \vee \overline{B}$ 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.

Throughout this paper, we assume the principle of semantic compositionality for natural language: the meaning of a complex expression is completely determined by the meaning of its constituents and the way they are combined. A corollary is the principle of substitution of equivalents: the meaning of a complex expression does not change when a constituent is replaced by another expression with the same meaning.

Under the compositionality assumption, the two views on the nature of meaning 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, and we expect that substituting one with the other in the context of a larger sentence should not affect its meaning. By contrast, if $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$ have different meanings, we expect them to make different semantic contributions when embedded in a larger sentence. In this case, substituting one with the other may well result in a different meaning, and possibly also in different truth conditions for the complex sentence.

To make this concrete, 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 $\neg(A \wedge B)$.

To investigate whether $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$ can indeed come apart in their truth values, we conducted a truth value judgment experiment based on the context described earlier. Our results show that while a vast majority of participants judged $\overline{A} \vee \overline{B} > \textsc{off}$ true in the given scenario, few judged $\neg(A \wedge B) > \textsc{off}$ true. This evidence favors a fine-grained view of meaning that teases apart $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$.[1]

[1]Contexts that make more fine-grained distinctions than those determined by truth conditions are often referred to as *hyperintensional* (Fox & Lappin 2005, McKay & Nelson 2014). Our experimental result can

In addition to providing empirical support for a fine-grained view on meaning, in this paper we develop a formal theory that explains the difference we observe between the counterfactuals $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$. Our theory builds on the framework of inquisitive semantics (Ciardelli, Groenendijk & Roelofsen to appear). In inquisitive semantics, the meaning of a sentence is not represented as a set of possible worlds, but as a set of such sets, referred to as the *alternatives* for the sentence. By representing sentence meaning as a set of alternatives, inquisitive semantics captures the intuition that a single sentence can evoke several possibilities. The difference between the two antecedents is captured by the fact that, while $\neg(A \wedge B)$ is associated with a single alternative, $\overline{A} \vee \overline{B}$ is associated with two distinct alternatives, corresponding to the two disjuncts.[2]

To explain how the presence of these different alternatives affect the truth conditions of counterfactuals, we combine inquisitive semantics with a proposal by Alonso-Ovalle (2006, 2009). According to this proposal, a counterfactual antecedent does not always contribute a unique assumption: when it is associated with multiple alternatives, as in the case of $\overline{A} \vee \overline{B}$, each of these alternatives is processed as a separate counterfactual assumption. This means that the consequent of $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$ ends up being assessed in different hypothetical situations in the two cases, resulting in different truth conditions for the two counterfactuals.

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.

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). 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

be viewed as showing that counterfactual antecedents are hyperintensional. Nevertheless, there is an important difference between counterfactual antecedents and other contexts which have been argued to be hyperintensional, such as belief attributions. In belief attribution claims, replacement of equivalent sentences is generally invalid: even *John believes it will rain within a fortnight* and *John believes it will rain within two weeks* may have different truth values. This is not the case for counterfactual antecedents: *If it had rained for a fortnight, the crops would have been ruined* has the same truth conditions as *If it had rained for two weeks, the crops would have been ruined.* Our proposal preserves the principle of substitution of equivalent antecedents, but it makes it harder for two sentences to count as equivalent.

[2]The semantic frameworks of alternative semantics (Hamblin 1973, Kratzer & Shimoyama 2002, Alonso-Ovalle 2006) and truth-maker semantics (Fine 2012a) are similar to inquisitive semantics in some crucial respects. However, some specific differences between these frameworks also matter for our concerns, as we discuss in Section 6.1. 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.

antecedents such as *If this match was struck, it would light.* In judging this sentence, we allow certain 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, and so on. 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 (Stalnaker 1968, 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''. Let us refer to a world where φ is true as a φ-world. 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 φ-worlds which are most similar to w.

In this paper we test the predictions of the minimal change requirement in a novel way. Crucially, we do not rely on any pre-defined assumption about similarity.[3] 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]

To make this concrete, consider our switches scenario. Suppose 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 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, that is, a most 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, this world must be an Off-world. Consequently, $\neg(A \wedge B) > \textsc{off}$ is predicted to be true.

Thus, regardless of what notion of world similarity we assume, ordering semantics predicts that in any context in which $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ are true, so is $\neg(A \wedge B) > \textsc{off}$. This means that 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 as formulated in Kratzer (1981a,b). We come back to this in Section 6.3.

[3] Many arguments against Lewis (1973) rely on such assumptions. For example, Fine (1975) argues that the minimal change requirement would wrongly predict *If Nixon had pressed the button there would have been a nuclear holocaust* to be false, because one can easily imagine a small change that prevents the button from working. These arguments rely crucially on specific assumptions about similarity (Lewis 1979)—in this case, the assumption that a world where a wire malfunction prevents the button from causing nuclear war counts as more similar to the actual world than one where nuclear war takes place.

[4] This way of arguing is in line with a methodological suggestion by Lewis (1979: p. 466f.): to evaluate whether ordering semantics is empirically accurate, we must determine the nature of the similarity ordering by reasoning backwards from the truth conditions of counterfactuals without imposing any *a priori* plausibility assumptions on the similarity ordering.

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.

In addition to presenting 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 binary distinction between facts that are at stake—or *foregrounded*—when making a counterfactual assumption and facts that are regarded as *backgrounded*. While backgrounded facts are held fixed in making the assumption, foregrounded facts can be manipulated 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 viewed as backgrounded; since backgrounded 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 backgrounded; 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. This does not allow us to reach any definite conclusion about the state of the light. Therefore, $\neg(A \wedge B) > \textsc{off}$ is not judged true.

Building on ideas from premise semantics (Kratzer 1981a,b) and causal reasoning (Pearl 2000), we develop a formal theory of counterfactuals that embodies these intuitions. Unlike Pearl (2000), our theory is not restricted to special kinds of antecedents but can deal with antecedents of arbitrary complexity. The ideas that we propose can also be used to extend Pearl's intervention-based account of counterfactuals to arbitrary antecedents.[5] 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 a factual background. Section 5 discusses additional patterns in our data, involving minority judgments and order effects, and suggests explanations based on our theory. Section 6 discusses the connections between our proposal and related work on counterfactuals. Section 7 concludes. A visual outline of the challenges we discuss and of the solutions we propose is given in Figure 2.

[5] For another proposal aiming to generalize the account of Pearl to arbitrary antecedents, see Briggs (2012).

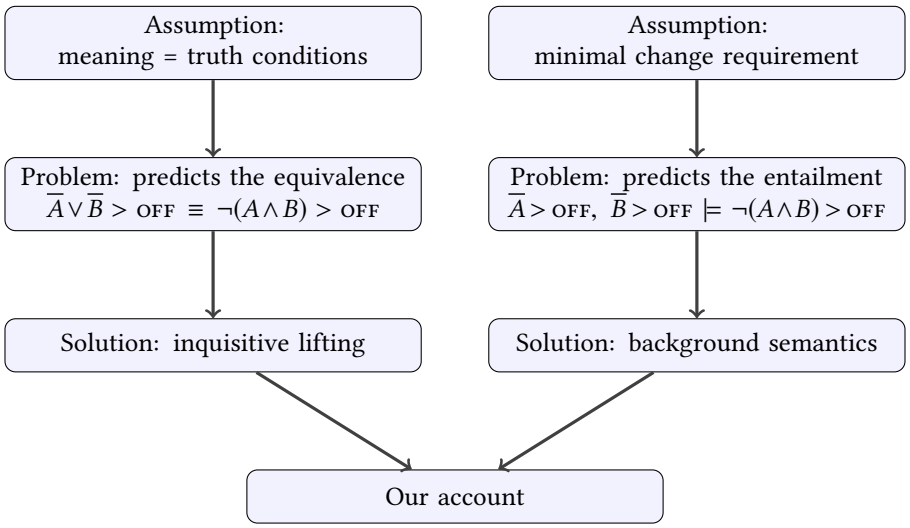


Figure 2: The paper at a glance.

2 Experiment

2.1 Hypotheses and predictions

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 $\neg(A \wedge B)$. As we discussed, assuming compositionality, the hypothesis that truth conditions completely determine meaning predicts that $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$ should have the same meaning as well, and should be judged in the same way in a given situation. By contrast, if meaning is not completely determined by truth conditions, then $\overline{A} \vee \overline{B}$ 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 $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$ 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 $\overline{A} > \textsc{off}$, $\overline{B} > \textsc{off}$, and $\neg(A \wedge B) > \textsc{off}$. As we argued in Section 1.2, the minimal change requirement predicts that in any context where $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ are both judged true, $\neg(A \wedge B) > \textsc{off}$ should be judged true as well. Thus, if $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ but not $\neg(A \wedge B) > \textsc{off}$ 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 (https://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.

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 all tests, incorrect responses to filler items accounted for the majority of rejected data.

All our experimental materials, instructions for participants, anonymized 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.

[6]Our figures are adapted from multiway switches © Colin M.L. Burnett (https://en.wikipedia.org/wiki/Multiway_switching#/media/File:3-way_switches_position_2.svg) CC BY-SA 3.0, via Wikimedia Commons.

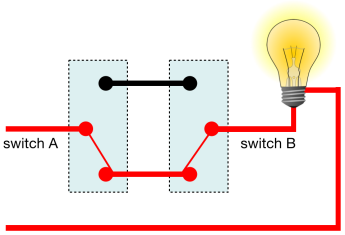


Figure 3: Pretest I. Switch A and B are both down, and the light is on.

2.3 Two pretests

2.3.1 Pretest I

Materials The goal of Pretest I was to confirm that the unembedded sentences $\overline{A} \vee \overline{B}$ 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 a wide majority of participants; we conclude that $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$ are both true in this scenario, which confirms that these sentences are truth-conditionally equivalent. This is in line with the experimental literature on disjunction since Paris (1973), which has generally found a preference for inclusive interpretations ("p or q or both") even in unembedded contexts. For example, Chevallier et al. (2008) find that the core meaning of *or* is inclusive, and that an exclusive interpretation ("p or q but not both") only arises when participants are forced to focus on the disjunction for at least three seconds.

The rate of acceptance was slightly higher for $\neg(A \wedge B)$ than for $\overline{A} \vee \overline{B}$. This difference was borderline significant ($\chi^2(2, N = 275) = 5.23, p = 0.07$). It could be attributed to noise, or it might indicate that a minority of participants interpreted *or* as exclusive disjunction. We come back to this point in Section 2.6.1.

2.3.2 Pretest II

Materials We assume that counterfactual sentences with simple antecedents (for example, *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,

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

$\overline{A}\vee\overline{B} > \textsc{off}$ 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 this 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.[7]

(3) 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 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 …

[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 of Veltman (2005). That discussion is not directly related to our main concerns here. The specific text in (3) 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.

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%

Our five target sentences, repeated from (1), are shown in (4). The labels that appear next to them were not shown to participants.

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

The word *both* in the antecedents of $\neg(A \wedge B) >$ OFF and $\neg(A \wedge B) >$ ON has been inserted to rule out an unwanted interpretation of the antecedent that is false or infelicitous in situations where the switches have different positions. That interpretation has been characterized as a homogeneity effect; for discussion, see Szabolcsi & Haddican (2004) and Magri (2014).

Our filler sentence is shown in (5). We ruled out data from those participants who failed to judged it false in the given context.

(5) 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 $\overline{A} >$ OFF, $\overline{B} >$ OFF, and $\overline{A} \vee \overline{B} >$ OFF, which were all judged true by a wide majority. In the second block, $\neg(A \wedge B) >$ OFF and $\neg(A \wedge B) >$ ON were generally judged false or indeterminate. The frequency difference between $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF is highly significant: $\chi^2(2, N = 734) = 197.84, p < 0.0001$. Differences within each block were not significant (first block:

[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 ($\overline{A} >$ OFF vs. $\overline{B} >$ OFF: $\chi^2(2, N = 491) = 0.90$; $\overline{A} >$ OFF vs. $\overline{A} \vee \overline{B} >$ OFF: $\chi^2(2, N = 618) = 0.28$; $\overline{B} >$ OFF vs. $\overline{A} \vee \overline{B} >$ OFF: $\chi^2(2, N = 597) = 0.37$; $\neg(A \wedge B) >$ OFF vs. $\neg(A \wedge B) >$ ON: $\chi^2(2, N = 572) = 0.38$).

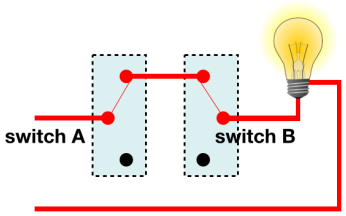


Figure 4: Post-hoc test I. There is no wire between the two "down" positions.

$\chi^2(4, N = 853) = 3.33, p = 0.5042$; second block: $\chi^2(2, N = 572) = 1.92, p = 0.3829$).

Crucially, our results show that $\overline{A} \vee \overline{B} > \text{OFF}$ and $\neg(A \wedge B) > \text{OFF}$ were judged differently, indicating that these two counterfactuals have different truth conditions. Moreover, while both $\overline{A} > \text{OFF}$ and $\overline{B} > \text{OFF}$ were judged true by a majority of participants, $\neg(A \wedge B) > \text{OFF}$ was not, 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, $\overline{A} > \text{OFF}$ and $\overline{B} > \text{OFF}$ are true, while $\neg(A \wedge B) > \text{OFF}$ is not, 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, $\overline{A} > \text{OFF}$, $\overline{B} > \text{OFF}$ and $\overline{A} \vee \overline{B} > \text{OFF}$, to the fourth, $\neg(A \wedge B) > \text{OFF}$.

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 difference in judgments between $\overline{A} \vee \overline{B} > \text{OFF}$ and $\neg(A \wedge B) > \text{OFF}$ observed 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 4, 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 (6):

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

If in our main experiment, the difference in truth value judgments between $\overline{A} \vee \overline{B} > \text{OFF}$ and $\neg(A \wedge B) > \text{OFF}$ is mainly due to context-independent factors, we expect to observe the same difference in this post-hoc test. If the difference tracks an actual difference in truth conditions, then in this new context, we expect that the pattern 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 this filler false.

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, the differences among the truth value judgments of the first four sentences were only marginally significant ($\chi^2(6, N = 346) = 11.26, p = 0.08$). Moreover, this time $\overline{A} >$ OFF, $\overline{B} >$ OFF, $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF were all judged true by a large majority of participants. This indicates that the difference in truth value judgments between $\overline{A} >$ OFF/$\overline{B} >$ OFF/$\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF that we observed in our main experiment is not driven by context-independent factors. More specifically, the dropoff in "true" judgments from any of the first four sentences to $\neg(A \wedge B) >$ OFF is always under 10% in this Post-hoc test, while in the main experiment, it is always over 40%. The vast discrepancy between these differences suggests that differences in processing load are responsible at most for a quarter of the dropoff observed in the main experiment.

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 test 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 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 ($F(1, 122) = 1.26$, $p = 0.26$). In any case, filtering out participants who answered the filler sentence wrong did not markedly affect the general

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%

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%

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.

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 judgments 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 $\neg(A \wedge B) >$

patterns we found in Post-hoc Tests II and III.

OFF in the main experiment is still significant: $\chi^2(2, N = 455) = 83.89, p < 0.0001$. Therefore, we can exclude the choice of auxiliary as a factor affecting our findings.[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 *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 (Horn 1985, Aloni 2016). However, theoretical considerations (e.g. Gazdar 1979, Chierchia 2004, Fox 2007, Spector 2007) suggest that implicature calculation or a silent exhaustivity operator might be responsible for what appears to be an exclusive interpretation of natural language disjunction in certain environments. One may therefore wonder whether an exclusive interpretation of *or* 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, was judged true by 81% of participants when both switches were down. In other words, *or* receives an exclusive interpretation in $\overline{A} \vee \overline{B}$ at most 19% of the time. By contrast, in our main experiment, $\overline{A} \vee \overline{B} >$ OFF was judged true almost 70% of the time, and $\neg(A \wedge B) >$ OFF only 22% of the time. The difference between these two proportions is so large (48%) that it is unlikely to be driven by exclusive interpretation of disjunctive antecedents. If this were the case, it would follow that

[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 we 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 sentences in the first block ($\overline{A} >$ OFF/$\overline{B} >$ OFF/$\overline{A} \vee \overline{B} >$ 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.

disjunctions receive exclusive interpretations in counterfactual antecedents at a much higher rate than in main clauses. The rate at which disjunctions receive exclusive interpretations is expected on theoretical grounds to be higher in main clauses than in contexts that license negative polarity items (e.g. Chierchia 2004). Such contexts include conditional antecedents as well as negation. Schwarz, Clifton & Frazier (2008) found that sentences like (7a) received exclusive interpretations 64.7% of the time and sentences like (7b) only 6.8% of the time.

(7) a. Maria asked Bob to invite Fred or Sam to the barbecue.
b. Maria asked Bob not to invite Fred or Sam to the barbecue.

For these reasons, we do not believe that the observed difference between $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$ is due primarily to exclusive interpretations of disjunction, and we will not pursue such an account.

A reviewer suggests that some participants might have interpreted the sentence $\overline{A} \vee \overline{B} > \textsc{off}$ with *or* taking wide scope over the counterfactual operator, resulting in the logical form $(\overline{A} > \textsc{off}) \vee (\overline{B} > \textsc{off})$. However, disjunctions in antecedents are not normally interpreted as having wide scope. For example, Alonso-Ovalle (2009) observes that even in scenarios where we accept (8b) as true, we still reject (8a) based on the falsity of (8c).[12]

(8) a. If we had had good weather or the sun had grown cold, we would have had a bumper crop.
b. If we had had good weather, we would have had a bumper crop.
c. If the sun had grown cold, we would have had a bumper crop.

It is also conceivable that in $\neg(A \wedge B) > \textsc{off}$, the antecedent *switch A and switch B were not both up* has been interpreted as meaning *switch A and switch B were both down*. This could happen in several ways: either by mistakenly reading *not both up* as *both not up*, or as a result of reading certain words as stressed, leading to narrow focus on these words. For example, some participants may have taken the word *up* to be focused and assumed that its salient alternative is *down*; others may have taken the word *both* to be focused and assumed that its only salient alternative is *neither*.[13] However, in all these cases we would expect a spike in 'true' judgments for $\neg(A \wedge B) > \textsc{on}$, which has the same antecedent; but only 21.5% of participants judged this sentence true. To further rule out these possibilities, we separately tested the unembedded sentence *Switch A and switch B are not both up* in a pictorial context that shows switch A up and switch B down. 76.9% of 290 participants judged it true in this scenario.

Finally, it is conceivable that the context in which our target sentences were interpreted varied systematically from one sentence to the next. We take this to be

[12]Experimentally, this suggestion could be assessed by testing our sentence $\overline{A} \vee \overline{B} > \textsc{off}$ in a context where only one of the switches controls the light, while the other has no effect. Participants who read the disjunction as having wide scope should still judge the sentence true in this scenario. A low rate of 'true' judgments would show that most participants take disjunction to have narrow scope.

[13]Participants who took *both* to be focused might also have interpreted it as giving rise to the alternative *just one*, rather than *neither*. This is unlikely to have happened often since these participants would be expected to interpret $\neg(A \wedge B) > \textsc{off}$ as true. As Table 3 shows, these participants numbered only 22.04% in our main experiment.

unlikely because we presented every target sentence in the same scenario and under the same conditions. One might worry that participants nevertheless have taken cues from the different antecedents of our target sentences to extrapolate systematically different contexts, which in turn could have given rise to different similarity orderings.

However, it is unclear what would have triggered the specific context shifts that would lead to our results. In particular, consider contexts that determine orderings in which toggling one switch represents less of a change than toggling two. Why should such contexts be systematically more available for some of our target sentences than for others? One might perhaps imagine that every time a sentence mentions a switch, this has a systematic pragmatic effect on its context; however, this hypothesis does not distinguish between $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$, since they both mention the same switches.[14]

2.6.3 Conclusion

Having excluded various confounds, we take the differences in native speakers' judgment on our sentences to track actual differences in truth value: under the most salient reading of these counterfactuals, $\overline{A} > \textsc{off}$, $\overline{B} > \textsc{off}$ and $\overline{A} \vee \overline{B} > \textsc{off}$ are true in our scenario, while $\neg(A \wedge B) > \textsc{off}$ and $\neg(A \wedge B) > \textsc{on}$ are not.[15]

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 conforms to the minimal change requirement.

With respect to the first question, we take our results to show that $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$ 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, under the compositionality assumption, we conclude that it is possible for two sentential clauses to have the same truth conditions and different meanings—which implies that meaning is not completely determined by truth conditions.

With respect to the second question, we take our results to show that $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ can be true in a context where $\neg(A \wedge B) > \textsc{off}$ is not true. As we discussed in

[14] The notion that antecedents systematically influence similarity orderings is conceptually problematic no matter if this influence is taken to be part of the pragmatics or part of the semantics. The pragmatic option assumes that the antecedent of a counterfactual can resolve contextual indeterminacy concerning the choice of the similarity ordering. If this possibility were invoked unrestrictedly and without a positive account of this pragmatic process, one could never assume that two distinct counterfactuals are interpreted in the same context. Thus, if antecedents could be responsible for systematic and unpredictable context shifts, it would become unclear how the theory may be empirically assessed at all. While proponents of ordering semantics do on occasion react to potential counterexamples by appealing to contextual indeterminacy (see for example Lewis 1979: 457f. Stalnaker 1981: 92f. and Stalnaker 1984: 130f.), to our knowledge a fully fledged pragmatic theory that both allows and suitably constrains context shifts has yet to be proposed for ordering semantics. As for the semantic option, it amounts to making the similarity ordering itself antecedent-relative. As Cross (2008) shows, this theory fails to validate even the weakest logic corresponding to ordering semantics, and can be seen to be an equivalent (but conceptually less parsimonious) formulation of the most general selection function framework, which just assumes a function f mapping each possible world w and antecedent φ to a set $f(w, \varphi)$ of φ-worlds. Cross concludes that such a theory is an ordering semantics in name only.

[15] We do not draw any conclusion at this stage as to whether $\neg(A \wedge B) > \textsc{off}$ and $\neg(A \wedge B) > \textsc{on}$ are false or simply lack a truth value in our context. We come back to this issue in Section 5.

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 assume.

3 Breaking de Morgan's law in conditional antecedents

In this section, we propose an explanation of the classically unexpected contrast between $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$. We saw that, assuming compositionality, such an explanation requires a notion of meaning which is more fine-grained than the truth-conditional one, as well as a theory of propositional connectives which invalidates de Morgan's law, teasing apart $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$. Our solution builds on the framework of inquisitive semantics (Ciardelli, Groenendijk & Roelofsen to appear), which supplies both these ingredients.[16]

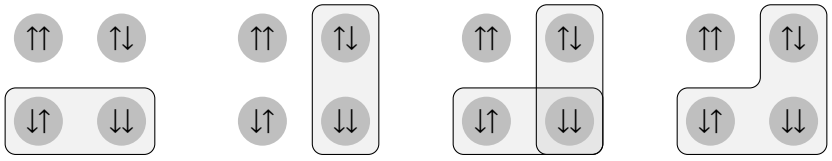
3.1 Inquisitive semantics

In inquisitive semantics, 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: $|\varphi| := \bigcup \mathrm{Alt}(\varphi)$. Thus, the inquisitive meaning of a sentence determines its truth conditions, but the converse is not 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 Figures 5(a) and 5(b).

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

[16] Obviously, inquisitive semantics is not the only approach that breaks de Morgan's law. One could, for example, base an explanation of our contrast on intuitionistic logic, where this law fails. However, as we will see in Section 3.1, inquisitive semantics has the merit of teasing apart $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$ while simultaneously accounting for their truth-conditional equivalence. Moreover, as we will see in Section 3.2, we can rely on a general recipe to transfer standard accounts of conditionals to the setting of inquisitive semantics. As far as we know, no such recipe is available for intuitionistic logic or other theories that break de Morgan's law.



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

Figure 5: The alternatives for our antecedents. ↑↑ represents a world where both switches are up, ↑↓ a world where A is up but B is down, etc.

Inquisitive semantics comes with a natural treatment of propositional connectives, obtained by associating these connectives with 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:[17]

(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 $\forall t \subseteq s:$ if $t \neq \emptyset$ then $t \not\models \varphi$

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, $\overline{A} \equiv \neg A$.[18]

(11) $\mathsf{Alt}(\neg A) = \{|\overline{A}|\}$

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 5(c).

(12) $\mathsf{Alt}(\overline{A} \vee \overline{B}) = \{|\overline{A}|, |\overline{B}|\}$

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 5(d).

(13) $\mathsf{Alt}(\neg(A \wedge B)) = \{\, W - |A \wedge B| \,\}$

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

[17] We build on the standard version of inquisitive semantics (see Ciardelli, Groenendijk & Roelofsen to appear), which has a solid logical foundation (Ciardelli & Roelofsen 2011, Ciardelli 2016b), and which has been assumed in most linguistic work based on the framework. Within the inquisitive semantics tradition, other accounts of connectives have also been investigated. In particular, *suppositional inquisitive semantics* (Groenendijk & Roelofsen 2015, Aher & Groenendijk 2015) treats negation differently from the standard system. In that system, however, de Morgan's laws are valid; therefore, it would not provide a suitable starting point for an account of our experimental findings.

[18] Recall that we are assuming that *up* and *down* are the only possible positions for our switches.

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.

3.2 Two assumptions for one antecedent

Having explained how the clauses $\overline{A} \vee \overline{B}$ and $\neg(A \wedge B)$ differ in meaning, the next step is to explain how this difference ends up affecting the truth conditions of the counterfactuals 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 counts as a distinct counterfactual assumption. In order for the 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 (2016a).[19] The starting point is an arbitrary truth-conditional account of counterfactuals, given in the form of a binary operation $\Rrightarrow$ (pronounced "then") which maps any two propositions p and q to a corresponding conditional proposition $p \Rrightarrow q$. Most existing accounts of counterfactuals, including ordering semantics (Stalnaker 1968, Lewis 1973), premise semantics (Kratzer 1981a), and some causal accounts (Kaufmann 2013, Santorio 2014, forthcoming[b]), can be seen as determining such a map.[20] The lifting recipe interprets a counterfactual sentence $\varphi > \psi$ by means of the following support clause.[21]

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 $s \subseteq (p \Rrightarrow q)$

If φ and ψ are non-inquisitive, that is, if $\mathsf{Alt}(\varphi) = \{|\varphi|\}$ and $\mathsf{Alt}(\psi) = \{|\psi|\}$, the clause yields a unique alternative for $\varphi > \psi$, which coincides with the counterfactual

[19] In Section 6.1 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 account for our experimental findings, but for reasons orthogonal to the central idea discussed here.

[20] In each of these accounts, the definition of the conditional proposition $p \Rrightarrow q$ 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) and Santorio (forthcoming[b]). However, the lifting recipe only needs access to the resulting operation on propositions—not to this additional structure.

[21] This clause is more general than what is needed for our immediate purposes. It is formulated with an eye towards consequents that provide two or more alternatives. We assume with Ciardelli (2016a) that this is the case for counterfactual questions such as *If switch A was down, would the light be on or off?* In our examples, this is not relevant, because all consequents provide a single alternative. More generally, we assume that declarative consequents always provide a single alternative. Following Ciardelli, Groenendijk & Roelofsen (to appear), we take this to be due to the presence of a silent declarative complementizer contributing a non-inquisitive closure operator, whose effect is to collapse multiple alternatives into one. We assume that this silent complementizer is prevented from appearing in antecedents by the presence of the complementizer *if*. The fact that consequents can have question syntax, but the complements of *if* cannot, has been taken as syntactic evidence that only consequents can be complementizer phrases (Iatridou 1991, Bhatt & Pancheva 2006).

proposition $|\varphi| \Rrightarrow |\psi|$ delivered by the given base account: $\mathrm{Alt}(\varphi > \psi) = \{\, |\varphi| \Rrightarrow |\psi| \,\}$.

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 interprets it as follows:

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

As in the previous cases, the counterfactual as a whole has a unique alternative, namely, the proposition $(|\overline{A}| \Rrightarrow |\textsc{Off}|) \cap (|\overline{B}| \Rrightarrow |\textsc{Off}|)$. However, this alternative is not the same proposition $|\overline{A} \vee \overline{B}| \Rrightarrow |\textsc{Off}|$ that would be delivered by the basic truth-conditional account. 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 equivalent to the conjunction of $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$.

This means that the truth conditions of our sentences will be correctly predicted 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 rendering $\neg(A \wedge B) > \textsc{off}$ true.

4 A background semantics for 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 context, $\overline{A} > \textsc{off}$ 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, Schulz 2011, Kaufmann 2013) incorporate the minimal change requirement in some form, and this leads them to predict that $\neg(A \wedge B) > \textsc{off}$ is true in any context in which $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ are true. Therefore, as a result of the minimal change requirement, these theories are not in a position to correctly predict our experimental findings, 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, in combination with the inquisitive lifting described in Section 3.2, explains our experimental results. For reasons that will become clear shortly, we refer to our theory as *background semantics*. 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 the literature on causal reasoning (Pearl 2000) and in causal versions of premise semantics (Schulz 2007, Kaufmann 2013). In Section 4.3 we

use these notions to formalize background semantics, and we show that this theory accounts for our experimental findings.

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 experimental results is the fact that the counterfactual $\neg(A \wedge B) > \textsc{off}$, repeated here as (14), is not judged true.

(14) 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.[22] 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 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 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 by an assumption a in case either of the following holds:

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

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 is a causal consequence of the assumption φ combined with the background facts.

This does not yet allow us to make any specific predictions, since different choices concerning what to put in the background may lead to different truth values. However, our experimental results can be explained if we assume a general preference for maximizing the set of backgrounded facts: that is, we assume that by default, the

[22] 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.

factual background consists of all and only the facts that are not called into question by the counterfactual assumption.[23]

Let us see how an account of the kind sketched here provides an explanation for our experimental findings. 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 judged true. Of course, the situation is completely analogous for the counterfactual $\overline{B} > \textsc{off}$.

As for the counterfactual $\overline{A} \vee \overline{B} > \textsc{off}$, according to our inquisitive account, it does not involve considering a single disjunctive assumption, but rather two distinct assumptions, namely, that A is down, and that B is down. Since on either of these assumption it follows that the light is off, $\overline{A} \vee \overline{B} > \textsc{off}$ is judged true.

Finally, 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 state of affairs 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. 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 transform this idea into a proper account, we first need to make a number of notions formally precise. To this we turn now.

4.2 The context for a causal account: causal models

Causal approaches to counterfactuals assume that the evaluation of a counterfactual takes place in the context of a network of causal relationships that allow for specific causal inferences. Formalizations of this idea within the framework of premise semantics have been proposed by Schulz (2007, 2011), Kaufmann (2013) and Santorio (2014, forthcoming[b]). 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 and a set of causal laws. Formally, a causal model over a set of possible worlds W is a pair $M = \langle V, L \rangle$ consisting of the following:

[23] Importantly, we propose to regard this only as a default choice, and not as an integral part of the semantics of counterfactuals. In Section 5 we discuss some evidence which points to the existence of non-maximal-background readings.

- A set V of *causal variables*, where a causal variable is a partition of the space of possible worlds. 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 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 $l = \langle C, E, m \rangle$ where C, the *cause set*, is a set of causal variables; E, the *effect*, is a causal variable not contained in C; and m, the *map*, is a function from settings of C to settings of E. Intuitively, m specifies how the value of the effect depends on the values of the causes. In line with Pearl (2000), we assume that each variable is the effect of at most one law in L.

  We say that a law $l = \langle C, E, m \rangle$ is obeyed at a world w if the value of E is in accordance with the law, that is, if $E_w = m(C_w)$. We call a world w law-abiding if it obeys all the laws in L.[24]

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. There is only one law; its cause set is $\{?A, ?B\}$, its effect is $?\textsc{On}$, and its map is:

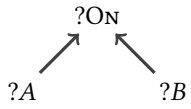
$$\{|A|, |B|\} \mapsto |\textsc{On}| \qquad \{|A|, |\overline{B}|\} \mapsto |\textsc{Off}|$$
$$\{|\overline{A}|, |\overline{B}|\} \mapsto |\textsc{On}| \qquad \{|\overline{A}|, |B|\} \mapsto |\textsc{Off}|$$

This law is obeyed 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.

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 $M = \langle V, L \rangle$, 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.[25] For instance, the causal graph of our example looks as follows:

[24] A refinement of this approach, inspired by Briggs (2012), would associate a potentially different causal model M_w to each possible world. For our present purposes we can assume the causal modal to be fixed, in line with other causal accounts.

[25] 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[b]) 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.



This graph shows that the variables $?A$ and $?B$ have a direct causal influence on the variable ?On, and that there are no other causal relations.

4.3 Formalizing background semantics

Let us now use the structure provided by a causal model to formulate precisely the account of counterfactuals outlined in Section 4.1. Under a maximal background interpretation, our proposal 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. Moreover, unlike Schulz (2011) and Briggs (2012), we interpret antecedents compositionally, that is, we operate on the propositions they denote rather than on the logical formulas that stand for them.

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 take the facts to be the values of the causal variables at the given world.

Definition 2 (Facts)**.**
The facts at a world w are the values of the causal variables at w. The set of facts at w is denoted $\mathcal{F}_w$. A fact Y_w is dependent on a fact X_w if X is an ancestor of Y in the causal graph of M.[26]

In our example, in the actual world we have three facts, corresponding to the true settings of our variables: $\mathcal{F}_w = \{|A|, |B|, |\text{On}|\}$. The fact $|\text{On}|$ is dependent on the facts $|A|$ and $|B|$, and no other causal dependencies hold. Our set of facts can be seen as the analogue of the *ordering source* in premise semantics (for more on the connection with premise semantics, see Section 6.3).

Next, to formulate our ideas we need to specify when a fact f contributes to the falsity of a proposition a at a world w. We take this to be the case if there is 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.

Definition 3 (Facts that contribute to the falsity of a proposition)**.**
A fact $f \in \mathcal{F}_w$ contributes to the falsity of a proposition a at w in case there exists some set of facts F such that F is consistent with a, but $F \cup \{f\}$ is not.[27]

[26]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 $a \Rrightarrow c$ 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, $\bigcap \mathcal{F}_w = \{w\}$. It is easy to verify that, given the account we are going to spell out, this assumption yields the above account of counterfactuals with true antecedents.

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

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

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

For an illustration, consider the proposition that A is down, $|\overline{A}|$, in our scenario. The unique maximal set of facts which is consistent with this proposition is $\{|B|, |\textsc{On}|\}$. Thus, the only fact that contributes to the falsity of $|\overline{A}|$ is $|A|$.

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 two distinct facts contribute to the falsity of $|\neg(A \wedge B)|$, namely, $|A|$ and $|B|$.

The next step is to stipulate which 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 contribute to its falsity, as well as anything which is dependent on these facts.

Definition 4 (Calling a fact into question)**.**
A proposition a calls into question a fact f at world w if either (i) f contributes to the falsity of a, or (ii) f is dependent on some fact which contributes to the falsity of a.

In our scenario, the assumption $|\overline{A}|$ calls into question the fact $|A|$, since this fact contributes to the falsity of $|\overline{A}|$, as well as the fact $|\textsc{On}|$, 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 depends on any other fact that does. As for the assumption $|\neg(A \wedge B)|$, it 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 $|\textsc{On}|$, 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 so far to constrain which 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 can be backgrounded.

Definition 5 (Backgrounds)**.**
A background for a proposition a at a world w is a set $\mathcal{B}(w, a) \subseteq \mathcal{F}_w$ of facts which are not called into question by a. A background map is a function $\mathcal{B}$ which maps each world w and proposition a to a corresponding background $\mathcal{B}(w, a)$.

For any assumption a and world w, we have a minimal and a maximal background: the minimal background is the empty set, while the maximal background, denoted $\mathcal{B}^{max}(w, a)$, is the set of all the facts which are not called into question by a at w. In our

scenario, the maximal 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: $\mathcal{B}^{max}(w, |\overline{A}|) = \{|B|\}$. On the other hand, the maximal background for the assumption that A and B are not both up, $|\neg(A \wedge B)|$, is the empty set, since we saw that all facts are called into question by this assumption: $\mathcal{B}^{max}(w, |\neg(A \wedge B)|) = \emptyset$.

To assess what follows from a given counterfactual assumption, we consider the hypothetical context created by the assumption. This is the set of those worlds in which the assumption is true, the background facts are held fixed, and the causal laws are obeyed.[28]

Definition 6 (Hypothetical context created by an assumption)**.**
Let $\mathcal{B}$ be a background map. The hypothetical context created by an assumption a at world w under $\mathcal{B}$, denoted $f_{\mathcal{B}}(w, a)$, is the set of those possible worlds where (i) a is true; (ii) all facts in $\mathcal{B}(w, a)$ are true; and (iii) all laws in L are obeyed.

A conditional proposition $a \Rrightarrow c$ is true under a background map $\mathcal{B}$ iff c is entailed by the hypothetical context created by a, that is, true everywhere in this context.[29]

Definition 7 (Truth conditions for counterfactuals)**.**
A conditional proposition $a \Rrightarrow c$ is true at a world w under a background map $\mathcal{B}$ just in case $f_{\mathcal{B}}(w, a) \subseteq c$. More formally, $a \Rrightarrow c = \{w \in W \mid f_{\mathcal{B}}(w, a) \subseteq c\}$.

Our account allows us to make specific predictions for the truth of counterfactuals only in combination with a particular background map. We now show that the truth-conditions reflected by our majority judgments are accounted for if we assume that the default strategy is to maximize the background, that is, to use the map $\mathcal{B}^{max}$. This means that, as a default, one retains all facts that are not directly or indirectly called into question by the counterfactual assumption.[30]

First consider the assumption that switch A is down, $|\overline{A}|$. We saw that the maximal background for this assumption is $\{|B|\}$. So, the hypothetical context created by the assumption consists of those law-abiding worlds where switch A is down and switch B is up. In all such worlds, the light is off. Therefore, under a maximal background interpretation, the proposition $|\overline{A}| \Rrightarrow |\textsc{Off}|$ is true. Since this proposition is the unique

[28] Under this definition, causal laws are always held fixed when making an assumption. Although sufficient for our purposes, this approach will not do in cases where the counterfactual assumption concerns variables which are causally dependent on others. In that case, we need to allow some of the laws to be discarded in order to make room for the assumption. This requires implementing a notion of *intervention* on the causal model in the style of Pearl (2000). Our theory can be adapted easily to handle simple counterfactuals: in this case, we just need to discard those laws whose effect contributes to the falsity of the assumption. However, nested counterfactuals such as the ones considered by Briggs (2012) require a more elaborate notion of intervention. We leave this extension for future work.

[29] Here we are only concerned with specifying under what conditions a counterfactual is true. We are not assuming that in all other circumstances the counterfactual is false. In some circumstances, it may fail to have a well-defined truth value, possibly due to the failure of a homogeneity presupposition to the effect that the consequent should have the same truth value across the hypothetical context (von Fintel 1997). We return to this issue in Section 5.

[30] In Section 5 we suggest that our minority judgments may arise from a different background choice, and we discuss what other factors, besides the given counterfactual assumption, may play a role in the determination of the background.

alternative that our inquisitive account assigns to the counterfactual $\overline{A} > \textsc{off}$, we correctly predict that this counterfactual is true.

Of course, the situation is analogous for the counterfactual $\overline{B} > \textsc{off}$. As for the counterfactual $\overline{A} \vee \overline{B} > \textsc{off}$, we saw 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.

Now consider the assumption that the switches are not both up, $|\neg(A \wedge B)|$. We saw that the maximal background for this assumption is empty. So, the hypothetical context created by the assumption just consists of those law-abiding worlds where the switches are not both up. This context includes worlds where only one switch is down and the light is off, as well as worlds where both switches are down and the light is on. Thus, neither $|\textsc{Off}|$ nor $|\textsc{On}|$ is entailed in this context, which means that neither $|\neg(A \wedge B)| \Rrightarrow |\textsc{Off}|$ nor $|\neg(A \wedge B)| \Rrightarrow |\textsc{On}|$ is true. According to our inquisitive account, the first proposition 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 in our scenario.[31]

Summing up, then, by combining the background semantics for 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. This is made possible by the combination of (i) an inquisitive account of conditionals, which is sensitive not only to the truth-conditions of the antecedent, but also to the alternatives that it introduces, and (ii) a procedure for making counterfactual assumptions which is not constrained by the minimal change requirement.

5 Explaining other aspects of our findings

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 background parameter in our theory.

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. However, our experimental results 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 $\overline{A} > \textsc{off}$, $\overline{B} > \textsc{off}$, and $\overline{A} \vee \overline{B} > \textsc{off}$ as indeterminate, rather than true (see Table 3).

[31] In this case, the prediction does not depend on the assumption of a maximal background interpretation: the empty set is the only background available for the assumption $|\neg(A \wedge B)|$.

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. 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 introspective judgments reported to us, it seems plausible that participants who judge $\overline{A} > \textsc{off}$, $\overline{B} > \textsc{off}$, and $\overline{A} \vee \overline{B} > \textsc{off}$ 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 and the causal law. 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 background parameter. Whereas the majority interpretation results from maximizing the background, the minority interpretation results from minimizing it—that is, from taking it to be the empty set (which always counts as a possible background). Under this background, our semantics indeed predicts that $a \Rrightarrow c$ is true in case c follows from a alone combined with the causal laws. In the scenario of our main experiment, the assumption that A is down together with the causal law does not lead to the conclusion that the light is off. Thus, under a purely causal interpretation, $\overline{A} > \textsc{off}$ is not predicted to be true, and similarly for $\overline{B} > \textsc{off}$ and $\overline{A} \vee \overline{B} > \textsc{off}$.

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 $\overline{A} > \textsc{off}$, $\overline{B} > \textsc{off}$, and $\overline{A} \vee \overline{B} > \textsc{off}$ 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 $\neg(A \wedge B) > \textsc{off}$ and to $\neg(A \wedge B) > \textsc{on}$ 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' judgment 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).[32]

5.2 Accounting for 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 shown 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 shown 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%).[33]

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, that is, 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 contextual factors beyond the given counterfactual assumption. In particular, if a previous sentence invites the 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.[34]

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 $|A|$ and $|B|$. In other words, to interpret $\overline{A} \wedge \overline{B} >$ OFF, one needs to attend to the possibility that the positions of the switches might both

[32]For cases in which a counterfactual antecedent introduces more than one assumption, it would also be natural to extend von Fintel's homogeneity presupposition to a presupposition to the effect that all the assumptions should lead to the same verdict about the consequent. This would predict that $\overline{A} \vee \overline{B} >$ OFF is neither true nor false in a scenario where only switch A controls the light, while B is inert. Such a presupposition is proposed and motivated by Santorio (forthcoming[a]) and Cariani & Goldstein (2017).

[33]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.

[34]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. Inspired by Warmbrōd (1981), von Fintel (2001) implements a similar idea: he takes counterfactuals to be strict conditionals over a set of worlds—the *modal horizon*—which expands as the discourse proceeds. Lin (2017) proposes a way to integrate this account with our background semantics by letting the modal horizon expand throughout the discourse to include the possible worlds selected by our hypothetical contexts.

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	42.49%	73	37.82%
$\neg(A \wedge B) >$ ON	102	21	20.59%	35	34.31%	46	45.10%

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%

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 foreground $|B|$, even though 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, which explains the 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.

6 Related work

In this section, we relate our work to relevant proposals on the semantics of counterfactuals. In Section 6.1 we compare our theory to other recent accounts which are similar to ours in that antecedents are not taken to provide a unique counterfactual assumption. In Sections 6.2 and 6.3 we discuss how background semantics relates, respectively, to the meta-linguistic approach to counterfactuals, and to the tradition of premise semantics. In Section 6.4 we discuss the issue of inferences from negated conjunctive antecedents, and we argue that these have a different origin than inferences

from disjunctive antecedents. We strengthen this point by considering connections between conditionals and modals.

6.1 Connections with other accounts of counterfactuals based on fine-grained meanings

Our account successfully teases apart the semantics of the two counterfactuals $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ 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 that make the semantics of conditionals sensitive to more than truth conditions. Proposals in this family include the theory of Fine (2012b), which is based on truth-maker semantics; the one of Willer (2015), which is based on a combination of dynamic semantics and inquisitive semantics; and the one of Alonso-Ovalle (2009), which is based on the framework of alternative semantics.

These accounts use a fine-grained representation of conditional antecedents mainly to validate the intuitive principle of simplification of disjunctive antecedents (SDA), while blocking full-fledged strengthening of the antecedent (AS), which is generally regarded as undesirable in conditional logic.

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

As Fine (1975) and Ellis, Jackson & Pargetter (1977) showed, this result is impossible to obtain for a compositional theory based on classical logic: for under these assumptions, SDA and AS are inter-derivable. This has been regarded as motivation for a fine-grained representation of antecedents. Our experimental results support the need for such a fine-grained representation: as we argued, a compositional account based only on truth conditions cannot explain the contrast we observed between $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF.

However, not all fine-grained accounts are in a position to account for this contrast. 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; for this reason, they still lead to the problematic prediction that $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF are equivalent. Thus, in spite of using a fine-grained semantics, these theories could not account for our findings without a revision of their underlying theories of propositional connectives.[35]

Now let us consider the theory of Alonso-Ovalle (2009). This theory is based on a treatment of disjunction in alternative semantics: each disjunct is taken to denote the

[35] 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 interesting ways. By assigning different meanings to the antecedents of $\overline{A} \vee \overline{B} >$ OFF and $\neg(A \wedge B) >$ OFF, this theory provides a suitable starting point for an account of our experimental results. 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.

singleton set of a proposition, and disjunction is taken to form the union of these sets, resulting in a two-element set. 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 is that disjunctive antecedents provide multiple assumptions. This insight 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 it in a different way, namely via the inquisitive lifting recipe developed in Ciardelli (2016a). This choice avoids two problems with Alonso-Ovalle's concrete proposal.

First, that proposal rests on alternative semantics, a framework which has not been equipped with a full-fledged theory of propositional connectives. In fact, Ciardelli, Roelofsen & Theiler (2016) argue that, in alternative semantics, it is difficult to provide a satisfactory treatment of conjunction; but such a treatment is of course required to represent the meanings of $\neg(A \wedge B) > \textsc{off}$ and $\neg(A \wedge B) > \textsc{on}$. Using the inquisitive lifting construction allows us to build on inquisitive semantics instead, a framework which comes with a well-developed theory of propositional connectives that shares many of the attractive features of classical logic (see also Ciardelli & Roelofsen 2011, Roelofsen 2013, Ciardelli & Roelofsen 2017).

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 also predicted to be true, contrary to our experimental findings. Thus, Alonso-Ovalle's theory cannot account for our experimental findings, 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 account of our experimental results.

6.2 Connections with the meta-linguistic theory

The background semantics we proposed in Section 4 is closely related to the so-called *meta-linguistic* theory of counterfactuals of Goodman (1947) (see also Chisholm (1946), Mackie (1962)). According to this theory, a counterfactual $A > C$ is true in case C can be inferred by causal laws from the antecedent A combined with certain true statements, the *co-tenables*, which can be held fixed while assuming A. Implementing this view requires a clear definition of the notion of co-tenability: which true statements are co-tenable with a given counterfactual assumption? Goodman considers a number of options, showing that each of them is either trivial, or leads to unacceptable predictions; in the end, he famously concludes that he is unable to provide a non-circular characterization of this notion.

Our background semantics can be seen as a concrete, formal implementation of the meta-linguistic theory: in background semantics, $A > C$ is true if C follows by causal laws from the antecedent A combined with certain true propositions which can be held fixed when making the assumption—the background facts. Thus, our proposal

can be seen as using causal models to provide a non-circular definition of co-tenability: a fact is co-tenable with an assumption A unless (i) it contributes to the falsity of A; or (ii) it is causally dependent on some fact that does; or (iii) the possibility of this fact being different is in some way salient in the context of evaluation (the case of non-maximal backgrounds).

6.3 Connections with premise semantics

Our background semantics also fits within the influential tradition of premise semantics (Kratzer 1981a,b). The first formulation of premise semantics is found in Kratzer (1981a); we will refer to it as *standard premise semantics*.[36] Subsequent accounts in the tradition of premise semantics differ from the standard formulation in various ways but tend to agree in outlook with it. Kratzer (1981a) articulates 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 premise 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 $g(w)$. In standard premise semantics, 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.[37]

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.

As Kratzer (1981a: fn. 8) notes, in order to validate commonly held inference patterns involving counterfactuals, one must assume that the ordering source contributed by the context of conversation stays the same during the inference. Given this assumption, standard premise semantics makes the same problematic prediction that we discussed in detail for ordering semantics. In fact, Lewis (1981) shows that standard premise semantics is equivalent to ordering semantics as defined in Lewis (1973) 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

[36] 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 background semantics and premise semantics also sets our theory apart from Veltman's.

[37] 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.

not rely on similarity being total, it follows that regardless of the particular ordering source that we consider, standard premise semantics still predicts that in any context where both $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ are true, $\neg(A \wedge B) > \textsc{off}$ is true as well, contrary to our experimental findings.

Broadly speaking, background semantics fits within the tradition of premise semantics. As in standard 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.

Nevertheless, there is a fundamental difference between the theory we propose and standard premise semantics. Our analysis departs from the basic idea laid out in Kratzer's quote, in that we do not incorporate the minimal change requirement. We propose that there is no general principle requiring us to add to the antecedent "as many facts as consistency permits". Rather, whenever we are faced with a counterfactual assumption, we determine a background of facts which are not at stake, and we hold all these facts fixed. While we do assume a preference for maximizing this background—and thus for avoiding gratuitous changes—this assumption is only a pragmatic default, and not part of the semantics of counterfactuals. More importantly, this is restricted to those facts that are not called into question by the counterfactual assumption; facts that *are* called into question are never backgrounded, even when doing so would not lead to inconsistency. This allows us to avoid the problematic prediction made by standard premise semantics and to explain why, in our context, the counterfactuals $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ are judged true, but $\neg(A \wedge B) > \textsc{off}$ is not.

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 a unique maximal set of premises consistent with a.

Among more recent systems that are formulated within the tradition of premise semantics, our system is similar to the ones of Kaufmann (2013) and Santorio (2014, forthcoming[b]), which like ours incorporate causal models in the style of Pearl (2000).

In the causal premise semantics of Kaufmann (2013), one does not, in general, add to the antecedent as many premises as consistency permits. For example, an assumption that switch A is down leads us to discard not only the fact that A is up, but also the causally dependent fact that the light is on, even if that fact could be added without violating consistency. This result is achieved by requiring all premise sets to be closed under causal ancestors. Nevertheless, the basic recipe for the interpretation of counterfactuals remains essentially the same as in standard premise semantics: one considers the maximal premise sets that are consistent with the antecedent and checks whether the consequent is entailed by each of them. For this reason, one can see Kaufmann's proposal as specifying how to use a causal structure to produce a suitable similarity relation on possible worlds. The interpretation of counterfactuals then proceeds in accordance with ordering semantics relative to the resulting model. This implies that the problematic entailment $\neg p > r,\ \neg q > r \models \neg(p \wedge q) > r$ is still valid in Kaufmann's theory.

In general, whenever a semantics can be seen as offering a criterion to determine a suitable similarity ordering, its logic is bound to include the standard conditional logic **P** (Kraus, Lehmann & Magidor 1990), the logic arising from the most general version of ordering semantics. This is because any entailment which is invalidated by the semantics can be falsified in a similarity-based model. In system **P**, the entailment $\neg p > r,\ \neg q > r \models \neg(p \wedge q) > r$ is valid. As we saw, however, this entailment is not valid in background semantics: in our scenario, the semantics predicts that $\neg A > \textsc{Off}$ and $\neg A > \textsc{Off}$ are true, but $\neg(A \wedge B) > \textsc{off}$ is not. This shows that background semantics does not validate system **P**, and therefore, that it cannot be seen as offering a procedure to determine a suitable similarity ordering. Rather, our semantics should be seen as departing from the similarity-based approach altogether. Investigating the logic arising from background semantics is an interesting task for future work.

Among premise semantic theories, the filtering semantics of Santorio (2014, forthcoming[b]) comes closest to our own. In this system, too, we do not look at the maximal sets of premises consistent with the given assumption; rather, the set of premises is "filtered" relative to a given antecedent, resulting in a single set of assumptions that plays roughly the same role as the combination of our background and the causal laws. Moreover, as in our account, filtering semantics invalidates the entailment $\neg p > r,\ \neg q > r \models \neg(p \wedge q) > r$ and, as a consequence, it should not be seen as a similarity-based theory. Nevertheless, filtering semantics in its existing form would wrongly predict $\neg(A \wedge B) > \textsc{off}$ to be true in our scenario. Background semantics can be seen as a proposal to fix this problem by adopting a different filtering procedure. In the absence of any further changes, the modified account would then make the wrong predictions about $\overline{A} \vee \overline{B} > \textsc{off}$, since the theory is compositional and based on classical logic. Therefore, to account for our experimental findings, the resulting theory still needs to be combined with a semantic theory that, like inquisitive semantics, teases apart the antecedents $\neg(A \wedge B)$ and $\overline{A} \vee \overline{B}$.

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 (15) does seem sound.

(15) 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 antecedent is inquisitive.[38] Nevertheless, since negative antecedents are *not* inquisitive, our theory does not explain the inference in (15) 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 apparent soundness of the inference in (15) separately. In this section, we show that our theory does account for this inference, by proving the following 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 at w, so is $\neg A > C$.

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

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

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

Proof of Proposition 2. It follows from Lemma 1 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 $\mathcal{B}^{max}(w, |\neg(A \wedge B)|) \subseteq \mathcal{B}^{max}(w, |\neg A|)$. 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 $\mathcal{B}(w, |\neg(A \wedge B)|) \subseteq \mathcal{B}(w, |\neg A|)$. This means that $f_{\mathcal{B}}(w, |\neg A|) \subseteq f_{\mathcal{B}}(w, |\neg(A \wedge B)|)$, that is, the hypothetical context created by $|\neg A|$ is included in the one created by $|\neg(A \wedge B)|$. Thus, if C is true everywhere in $f_{\mathcal{B}}(w, |\neg(A \wedge B)|)$, it is also true everywhere in $f_{\mathcal{B}}(w, |\neg A|)$.[39] □

Thus, provided that the propositions that Nixon resigned and that Agnew resigned are facts in our causal model, the soundness of (15) is indeed predicted on our account.

However, in our account the validity of SNCA has a different origin than the one of SDA: SDA stems from the presence of multiple alternatives in the antecedent, which provide multiple counterfactual assumptions; by contrast, SNCA stems from the specific workings of our procedure for making counterfactual assumptions.

Independent evidence for the fact that SDA and SNCA have different origins comes from looking at inferences involving modals. There is wide agreement in the literature that SDA inferences such as (16) are related to free choice inferences under modal operators, illustrated by (17).

(16) a. If Mr. X wore a top hat or a fedora, he would blend in with the crowd.
b. So, if he wore a top hat, he would blend in with the crowd.

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

[39] This proof extends to an arbitrary background function $\mathcal{B}$ under the assumption that, if a calls into question everything that a' does, then any fact that is foregrounded when making the assumption a is also foregrounded when making the assumption a' (that is, $\mathcal{B}(w, a') \subseteq \mathcal{B}(w, a)$). Moreover, the proof extends to the case in which we allow laws to be discarded in the way suggested in Footnote 28.

(17) 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, as 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).

If free choice inferences are indeed 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 true: while 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 $\neg(A \wedge B) > C$, such as (18a), and modal sentences $\Diamond\neg(A \wedge B)$, such as (19a).

(18) a. If Mary had not spoken both Arabic and Bengali, she wouldn't have been hired.
b. So, if she had not spoken Arabic, she wouldn't have been hired.

(19) a. Mary might not speak both Arabic and Bengali.
b. # So, she might not speak Arabic.

The counterpart of the inference in (18) is not valid in the modal setting in (19). Consider the surface scope readings of these sentences (that is, $\Diamond\neg(A \wedge B)$ for (19a) and $\Diamond\neg A$ for (19b)): one may doubt whether Mary can speak both Arabic and Bengali, yet know for a fact that she does speak Arabic. In that situation, the inference in (19b) seems unwarranted. In fact, right after (19a), the speaker may follow up with an emphatic "but she certainly does speak Arabic!".

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

7 Conclusion

In this paper we have 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 experimental results 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 results 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 experimental results 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 (2016a), 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 findings cannot be accounted for. To put it differently, on a theory that implements the minimal change requirement, the majority judgments that we found are predicted to be jointly 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 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 combined with inquisitive semantics, predicts our majority judgments and explains various other patterns in our experimental results.

Appendix

A 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 f 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 proposi-

tional connectives are concerned, we have $|\neg A| = \overline{|A|}$ and $|\neg(A \wedge B)| = \overline{|A| \cap |B|}$. 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 contributes to the falsity of $\overline{f}$ at w is f;
2. the facts that contribute to the falsity of $\overline{f \cap g}$ 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.[40] 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 $\overline{f}$.
   - 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 contribute to 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, $H \subseteq G$.

[40] 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.

Acknowledgments

Champollion, Ciardelli & Zhang (2016) is an earlier version of the first part of this paper. For comments and discussion, we thank Luis Alonso-Ovalle, Rebekah Baglini, Justin Bledin, Joseph DeVeaugh-Geiss, Kit Fine, Ethan Jerzak, Angelika Kratzer, Dan Lassiter, Johannes Marti, Robert van Rooij, Paolo Santorio, Katrin Schulz, Anna Szabolcsi, Frank Veltman, Malte Willer, and audiences at NYU, SALT 26, the Fourth Workshop on Natural Language and Computer Science (NLCS 2016), the Workshop on Logic and Algorithms in Computational Linguistics 2017, the *Philosophy meets Linguistics* workshop in Zürich, the InqBnB1 workshop in Broek in Waterland, LENLS 14, and in Utrecht, Paris, Göttingen, Harvard University, and Fudan University (Shanghai), as well as two anonymous reviewers and the editor, Stefan Kaufmann. Special thanks to Floris Roelofsen. Ivano Ciardelli's research was financially supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 680220). 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. https://doi.org/10.1007/s11050-007-9010-2.

Aloni, Maria. 2016. Disjunction. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, Winter 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/disjunction/.

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

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

Bhatt, Rajesh & Roumyana Pancheva. 2006. Conditionals. In Martin Everaert & Henk van Riemsdijk (eds.), *The Blackwell companion to syntax*, 638–687. Blackwell Publishing. https://doi.org/10.1002/9780470996591.ch16.

Briggs, Rachael. 2012. Interventionist counterfactuals. *Philosophical Studies* 160(1). 139–166. https://doi.org/10.1007/s11098-012-9908-5.

Cariani, Fabrizio & Simon Goldstein. 2017. Conditional heresies. https://philarchive.org/archive/CARCH-5.

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

Chevallier, Coralie, Ira A. Noveck, Tatjana Nazir, Lewis Bott, Valentina Lanzetti & Dan Sperber. 2008. Making disjunctions exclusive. *The Quarterly Journal of Experimental Psychology* 61(11). 1741–1760. https://doi.org/10.1080/17470210701712960.

Chierchia, Gennaro. 2004. Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In Adriana Belletti (ed.), *Structures and beyond*, vol. 3 (The Cartography of Syntactic Structures), 39–103. Oxford University Press.

Chisholm, Roderick M. 1946. The contrary-to-fact conditional. *Mind* 55(220). 289–307. http://www.jstor.org/stable/2250757.

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

Ciardelli, Ivano. 2016b. *Questions in logic*. University of Amsterdam dissertation. http://hdl.handle.net/11245/1.518411.

Ciardelli, Ivano, Jeroen Groenendijk & Floris Roelofsen. to appear. *Inquisitive semantics*. Oxford University Press.

Ciardelli, Ivano & Floris Roelofsen. 2011. Inquisitive logic. *Journal of Philosophical Logic* 40(1). 55–94. https://doi.org/10.1007/s10992-010-9142-6.
Ciardelli, Ivano & Floris Roelofsen. 2017. Hurford's constraint, the semantics of disjunctions, and the nature of alternatives. *Natural Language Semantics* 25(3). 199–222. https://doi.org/10.1007/s11050-017-9134-y.
Ciardelli, Ivano, Floris Roelofsen & Nadine Theiler. 2016. Composing alternatives. *Linguistics and Philosophy*. 1–36. https://doi.org/10.1007/s10988-016-9195-2.
Cross, Charles. 2008. Antecedent-relative comparative world similarity. *Journal of Philosophical Logic* 37(2). 101–120. https://doi.org/10.1007/s10992-007-9061-3.
Ellis, Brian, Frank Jackson & Robert Pargetter. 1977. An objection to possible-world semantics for counterfactual logics. *Journal of Philosophical Logic* 6(1). 355–357. https://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. https://doi.org/10.1007/s11049-015-9305-9.
Fine, Kit. 1975. Critical notice. *Mind* 84(335). 451–458. https://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. https://doi.org/10.1007/s11229-012-0094-y.
Fine, Kit. 2012b. Counterfactuals without possible worlds. *The Journal of Philosophy* 109(3). 221–246. https://doi.org/10.5840/jphil201210938.
Fine, Kit. 2014. Truth-maker semantics for intuitionistic logic. *Journal of Philosophical Logic* 43(2-3). 549–577. https://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. https://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, chap. 3, 123–152. Cambridge, MA: MIT Press. http://mit.edu/fintel/fintel-2001-counterfactuals.pdf.
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. http://mit.edu/fintel/fintel-2004-kof.pdf.
Fox, Chris & Shalom Lappin. 2005. *Foundations of intensional semantics*. Oxford, UK: Blackwell Publishing. https://doi.org/10.1002/9780470773543.
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. https://doi.org/10.1057/9780230210752_4.
Gazdar, Gerald. 1979. *Pragmatics: Implicature, presupposition and logical form*. New York, NY: Academic Press.
Goodman, Nelson. 1947. The problem of counterfactual conditionals. *The Journal of Philosophy* 44(5). 113–128. https://doi.org/10.2307/2019988.
Groenendijk, Jeroen & Floris Roelofsen. 2015. Towards a suppositional inquisitive semantics. In Martin Aher, Daniel Hole, Emil Jeřábek & Clemens Kupke (eds.), *Logic, language, and computation: 10th international Tbilisi symposium (TbiLLC*

2013), 137–156. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-46906-4_9.

Groenendijk, Jeroen & Martin Stokhof. 1990. Dynamic Montague grammar. In László Kálmán & László Pólos (eds.), *2nd symposium on logic and language*, 3–48. Budapest, Hungary: Eötvös Loránd Press. http://hdl.handle.net/11245/1.428383.

Halpern, Joseph Y. 2013. From causal models to counterfactual structures. *The Review of Symbolic Logic* 6(2). 305–322. https://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 dissertation. http://semanticsarchive.net/Archive/jA2YTJmN.

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

Horn, Laurence R. 1985. Metalinguistic negation and pragmatic ambiguity. *Language* 61(1). 121–174. https://doi.org/10.2307/413423.

Iatridou, Sabine. 1991. *Topics in conditionals*. Massachusetts Institute of Technology dissertation. http://hdl.handle.net/1721.1/13521.

Kamp, Hans. 1973. Free choice permission. *Proceedings of the Aristotelian Society* 74(1). 57–74. https://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. https://doi.org/10.1002/9780470758335.ch8.

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

Kratzer, Angelika. 1981a. Partition and revision: the semantics of counterfactuals. *Journal of Philosophical Logic* 10(2). 201–216. https://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. https://doi.org/10.1002/9780470758335.ch12.

Kratzer, Angelika & Junko Shimoyama. 2002. Indeterminate pronouns: The view from Japanese. In Yukio Otsu (ed.), *3rd Tokyo conference on psycholinguistics*, 1–25. https://doi.org/10.1007/978-3-319-10106-4_7.

Kraus, Sarit, Daniel Lehmann & Menachem Magidor. 1990. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial intelligence* 44(1-2). 167–207. https://doi.org/10.1016/0004-3702(90)90101-5.

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

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

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

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

Lin, Shih-Yueh Jeff. 2017. Eliminating similarity in dynamic approaches to counterfactuals. NYU manuscript.
Mackie, John Leslie. 1962. Counterfactuals and causal laws. In Ronald Joseph Butler (ed.), *Analytical philosophy*, vol. 1, 66–80. Oxford, UK: Basil Blackwell.
Magri, Giorgio. 2014. An account for the homogeneity effects triggered by plural definites and conjunction based on double strengthening. In Salvatore Pistoia Reda (ed.), *Pragmatics, semantics and the case of scalar implicatures* (Palgrave Studies in Pragmatics, Language and Cognition), 99–145. London, UK: Palgrave Macmillan. https://doi.org/10.1057/9781137333285_5.
McKay, Thomas & Michael Nelson. 2014. Propositional attitude reports. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2014/entries/prop-attitude-reports/.
Nute, Donald. 1975. Counterfactuals and the similarity of words. *The Journal of Philosophy* 72(21). 773–778. https://doi.org/10.2307/2025340.
Nute, Donald. 1980. Conversational scorekeeping and conditionals. *Journal of Philosophical Logic* 9(2). 153–166. https://doi.org/10.1007/bf00247746.
Paris, Scott G. 1973. Comprehension of language connectives and propositional logical relationships. *Journal of Experimental Child Psychology* 16(2). 278–291. https://doi.org/10.1016/0022-0965(73)90167-7.
Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/cbo9780511803161.
Rawlins, Kyle. 2013. (Un)conditionals. *Natural Language Semantics* 21(2). 111–178. https://doi.org/10.1007/s11050-012-9087-0.
Roelofsen, Floris. 2013. Algebraic foundations for the semantic treatment of inquisitive content. *Synthese* 190(1). 79–102. https://doi.org/10.1007/s11229-013-0282-4.
van Rooij, Robert. 2006. Free choice counterfactual donkeys. *Journal of Semantics* 23(4). 383–402. https://doi.org/10.1093/jos/ffl004.
van Rooij, Robert. 2008. Towards a uniform analysis of *any*. *Natural Language Semantics* 16(4). 297–315. https://doi.org/10.1007/s11050-008-9035-1.
Santorio, Paolo. Forthcoming(a). Alternatives and truthmakers in conditional semantics. *The Journal of Philosophy*.
Santorio, Paolo. Forthcoming(b). Interventions in premise semantics. *Philosophers' Imprint*.
Santorio, Paolo. 2014. Filtering semantics for counterfactuals: bridging causal models and premise semantics. *24th Semantics and Linguistic Theory Conference (SALT 24)*. 494–513. https://doi.org/10.3765/salt.v24i0.2430.
Schulz, Katrin. 2007. *Minimal models in semantics and pragmatics: Free choice, exhaustivity, and conditionals*. Amsterdam, Netherlands: University of Amsterdam dissertation. 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. https://doi.org/10.1007/s11229-010-9780-9.
Schwarz, Florian, Charles Clifton & Lyn Frazier. 2008. Strengthening 'or': effects of focus and downward entailing contexts on scalar implicatures. In Jan Anderssen, Keir Moulton, Florian Schwarz & Cherlon Ussery (eds.), *Semantics and processing*,

vol. 39 (University of Massachusetts Occasional Papers in Linguistics). Amherst, MA: Graduate Linguistic Student Association.

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

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. https://doi.org/10.1057/9780230210752_9.

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

Stalnaker, Robert C. 1981. A defense of conditional excluded middle. In William L. Harper, Robert Stalnaker & Glenn Pearce (eds.), *Ifs: conditionals, belief, decision, chance and time*, 87–104. Dordrecht, Netherlands: Springer Netherlands. https://doi.org/10.1007/978-94-009-9117-0_4.

Stalnaker, Robert C. 1984. *Inquiry*. Cambridge, MA: MIT Press.

Szabolcsi, Anna & Bill Haddican. 2004. Conjunction meets negation: a study in cross-linguistic variation. *Journal of Semantics* 21(3). 219–249. https://doi.org/10.1093/jos/21.3.219.

Veltman, Frank. 1976. Prejudices, presuppositions, and the theory of counterfactuals. In *Amsterdam papers in formal grammar. 1st Amsterdam colloquium*, 248–282. Amsterdam, Netherlands: University of Amsterdam. http://hdl.handle.net/11245/1.428635.

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

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

Willer, Malte. 2015. Simplifying counterfactuals. In Thomas Brochhagen, Floris Roelofsen & Nadine Theiler (eds.), *20th Amsterdam colloquium*, 428–437. Amsterdam, Netherlands: ILLC Publications. http://philosophy.uchicago.edu/faculty/files/willer/Simplifying%20Counterfactuals.pdf.

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.