Breaking de Morgan's law in counterfactual antecedents

Lucas Champollion
New York University
champollion@nyu.edu

Ivano Ciardelli
University of Amsterdam
i.a.ciardelli@uva.nl

Linmin Zhang
New York University
linmin.zhang@nyu.edu

To be presented at *Semantics and Linguistic Theory 26*, May 12-15, 2016

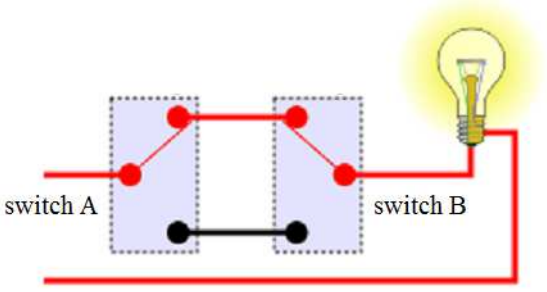
Abstract

We present the results of a web survey indicating that (ia) can be true in situations where (ib) is false.

(i) a. If switch A was down or switch B was down, the light would be off.
b. If switch A and switch B were not both up, the light would be off.

Assuming that the antecedents of these sentences are correctly analyzed as $(\neg A \vee \neg B)$ and as $\neg(A \wedge B)$, this challenges any compositional account of counterfactuals that interprets the two antecedents in classical logic: by de Morgan's law, their denotations are identical. We show that one can distinguish between (ia) and (ib) on a principled basis by interpreting their antecedents in inquisitive logic.

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 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 . . .*



We identify and meet novel challenges for theories of counterfactuals based on a Mechanical Turk survey. Each participant saw the figure and text above (based on Lifschitz, 1990) and two sentences in random order: one of the targets in (1), and the filler in (2), which is false.

(1) a. If switch A was down, the light would be off.
b. If switch B was down, the light would be off.
c. If switch A or switch B was down, the light would be off.

d. If switch A and switch B were not both up, the light would be off.
e. If switch A and switch B were not both up, the light would be on.

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

We implemented surveys via Turktools (Erlewine and Kotek, 2016). In a pre-test, all these sentences were judged natural (5 or higher on a 7 point scale), which we take to suggest that none of them is likely to be confusing or ungrammatical. In the survey, we asked participants to judge each sentence as *true, false* or *indeterminate*. We reduced noise by discarding results from participants who did not judge the filler sentence false, or who were not native speakers of American English. The remaining responses show a striking pattern:

Ex.	Translation	N	True	(%)	False	(%)	Indet.	(%)
(1a)	$\neg A > \textsc{off}$	255	169	**66.3%**	6	2.4%	80	31.4%
(1b)	$\neg B > \textsc{off}$	234	153	**65.4%**	7	3.0%	74	31.6%
(1c)	$(\neg A \vee \neg B) > \textsc{off}$	346	242	**69.9%**	12	3.5%	92	26.6%
(1d)	$\neg(A \wedge B) > \textsc{off}$	356	80	**22.5%**	129	36.2%	147	41.3%
(1e)	$\neg(A \wedge B) > \textsc{on}$	200	43	**21.5%**	63	31.5%	94	47.0%

Our results fall naturally into two blocks, as indicated by the dashed line. Differences across blocks were highly significant on a χ^2 test ($p < 0.0001$); differences within them were not (1^{st} block: $p = 0.36$ and higher; 2^{nd} block: $p = 0.43$). Here, we focus on the task of accounting for the column in bold, or in other words, explaining when counterfactuals are judged true.

The conjunction problem. The contrast between (1a)/(1b) and (1d) is incompatible with traditional minimal-change semantics of counterfactuals, according to which a counterfactual $\varphi > \psi$ is true (roughly) whenever every closest φ-world is a ψ-world (Lewis, 1973). In such a semantics, the fact that (1d) is not true requires that some closest $\neg(A \wedge B)$-world w fail to be an OFF-world; now, w must either be a closest $\neg A$-world or a closest $\neg B$-world: otherwise, some $\neg(A \wedge B)$-world would be closer than w to the actual world. In the first case, not every closest $\neg A$-world is an OFF-world, so (1a) is not predicted true, contrary to fact; in the second case, the same problem arises for (1b).

The de Morgan's Law problem. The contrast between (1c) and (1d) challenges any fully compositional theory of counterfactuals that builds on classical logic. The antecedents of (1c) and (1d) are classically equivalent by de Morgan's law: $\neg A \vee \neg B \equiv \neg(A \wedge B)$. As such, they are assigned identical truth-conditions. But then, by compositionality, this identity should propagate to the whole conditional. In fact, our results even challenge some theories built on less familiar logics, such as Fine (2012), which still validates de Morgan's law.

Breaking de Morgan's law. The difference that we observe between (1c) and (1d) finds a natural explanation once we move from a purely truth-conditional notion of meaning to a finer-grained one, such as that provided by the framework of *inquisitive semantics* (Ciardelli *et al.*, 2013). In inquisitive semantics, the meaning of a sentence φ is captured not by a single proposition, but rather by a set $\mathsf{Alt}(\varphi)$ of propositions, called the alternatives for φ. The set $|\varphi|$ of worlds where φ is true, called the *truth-set* of φ, is defined as the union of these alternatives: $|\varphi| = \bigcup \mathsf{Alt}(\varphi)$. Although the antecedents of (1c) and (1d) are true at the same set of worlds, in accordance with classical logic, their meaning consists of different alternatives. For the disjunctive antecedent in (1c), inquisitive semantics yields two alternatives: $\mathsf{Alt}(\neg A \vee \neg B) = \{|\neg A|, |\neg B|\}$. The antecedent of (1d), on the other hand, has

a unique alternative, which coincides with its truth-set: $\mathsf{Alt}(\neg(A \wedge B)) = \{|\neg(A \wedge B)|\}$. In the spirit of Alonso-Ovalle (2009), we can now assume that when an antecedent provides multiple alternatives, each of them is treated by the conditional as a separate assumption.[1] In order for a counterfactual to be true, the consequent must follow on each assumption. This ensures that (1c) is interpreted in effect as the conjunction of (1a) and (1b), and differently from (1d). This explains the strong similarity between the response pattern of (1c) and those of (1a) and (1b), and paves the way towards explaining the difference between (1c) and (1d).

Our baseline: a causal account. We are now left with the conjunction problem: predicting that (1a) and (1b) are true but (1d) and (1e) are not. We follow Kaufmann (2013) in adopting the causal approach of Pearl (2000); such an approach has been argued to be well suited for (1a)/(1b) (Schulz, 2007). We now present Kaufmann's system in simplified form; the modifications we will propose to account for (1d)/(1e) scale up to his full framework, and we inherit its strengths. We assume a causal structure and a set of laws. The causal structure is a directed acyclic graph $\langle V, E\rangle$ whose vertices, the *variables*, are partitions over the set of possible worlds; its edges E encode the direction of causal influence. The laws describe the correlations between the variables in the causal structure. In our example, that structure is given by the graph $?a \longrightarrow ?l \longleftarrow ?b$ whose variables $?a$, $?b$, and $?l$ correspond to the state of the two switches and the light, respectively. The *settings* of the variable $?a$ are the propositions "switch A is up", denoted a, and "switch A is down", denoted $\overline{a}$. Similarly, the settings of $?b$ are the propositions b and $\overline{b}$, and those of $?l$ are the propositions l and $\overline{l}$. The laws encode the behavior of the circuit. To simplify the presentation, we represent them compactly by the proposition $l \leftrightarrow (a \leftrightarrow b)$.

Given a causal structure $\langle V, E\rangle$ and a world w, typically the actual world, a *premise set* is a set S of settings of variables in V such that: (i) each element of S is true at w; and (ii) if S contains a setting of a variable v, it also contains a setting of of each ancestor of v. In our case, this means no premise set can contain l without also containing both a and b.

In our case, the premise sets are $\varnothing, \{a\}, \{b\}, \{a, b\}$, and $\{a, b, l\}$. To evaluate a counterfactual, among those premise sets that are consistent with the counterfactual assumption we choose the maximal ones, which we will call *backdrops*. Intuitively, backdrops hold remote worlds at bay. Requiring backdrops to be maximal is analogous to requiring worlds to be maximally close in Lewis (1973). For example, the unique alternative for the antecedent of (1a), $|\neg A| = \overline{a}$, is consistent with the premise sets $\varnothing$ and $\{b\}$; so $\{b\}$ is the only backdrop of $|\neg A|$. For law-abiding antecedents, Kaufmann's proposal boils down to the following:

(3) $p > q$ is true iff for each backdrop B of p, the intersection of (i) p, (ii) the propositions in B, and (iii) the laws entails q.

Consider the predictions of this recipe for (1a) and (1b). The antecedent of (1a) says that switch A is down; its backdrop, $\{b\}$, says that switch B is still up. Together with the laws, this entails that the light is off. Thus, (1a) is correctly predicted true. The situation is similar for (1b). However, this recipe does not make the right predictions for (1d): $\neg(A \wedge B)$ has two backdrops: $\{a\}$ and $\{b\}$. Each of them together with the antecedent entails that exactly one switch is up. Given the laws, in each case this means the light is off, which predicts (1d) to be true. In this case, we are retaining too much of the actual state of affairs.

[1] Our account departs from Alonso-Ovalle's in two ways: we do not build on Lewis (1973), and we adopt inquisitive rather than alternative semantics. We will discuss our reasons for these departures in the talk.

Adding grounds to the account. To solve the conjunction problem, we propose that in making a counterfactual assumption, one considers each "way" in which that assumption may obtain in the scenario. This is what forces attention to the case where both switches are down in (1d). To formalize this idea, we introduce the notion of *grounds*.

Let V be a set of variables, typically from a causal structure, and let p be a proposition. A *setting* of V is a set S that contains one setting for each variable in V; we call S a *p-setting* if $\bigcap S \subseteq p$, and a *$\overline{p}$-setting* if $\bigcap S \subseteq \overline{p}$, where $\overline{p}$ is the complement of p. We say that the set of variables V *controls* the proposition p if each of its settings is either a p-setting or a $\overline{p}$-setting. A *ground for p* is a p-setting of a minimal set of variables that controls p.

Now, take a causal structure $\langle V, E\rangle$, a world w and a ground G. A *backdrop of G* is a set that is maximal among those premise sets that are consistent with the intersection of all the propositions in G. We amend (3) as follows:

(4) $p > q$ is true iff for each ground G of p, for each backdrop B of G, the intersection of (i) the propositions in G, (ii) those in B, and (iii) the laws entails q.

Simple counterfactuals like (1a) are not affected by this move. For, the unique alternative for the antecedent of (1a) is $|\neg A| = \overline{a}$; this proposition is minimally controlled by the set $\{?a\}$ and has just one ground, $\{\overline{a}\}$. So, we still predict (1a) (and (1b)) to be true; and, given our account of antecedents with multiple alternatives, we also still predict (1c) to be true.

However, now we correctly predict that (1d) and (1e) are not true. Their antecedent has just one alternative, the proposition $|\neg(A \wedge B)|$, which has three grounds: $\{a, \overline{b}\}$, $\{\overline{a}, b\}$, and $\{\overline{a}, \overline{b}\}$. Given the laws, the first and second ground entail that the light is off, but the third entails that it is on. Hence, no firm conclusion can be drawn about whether the light would be on or off, in line with the high proportion of indeterminate answers to (1d) and (1e).

Accounting for the three-way split. The recipe in (4) merely gives conditions for truth, and does not by itself predict that (1d)/(1e) should be judged false. It may be complemented with an account of presupposition that allows us to categorize non-true sentences as either *false* or *indeterminate*. For example, those speakers that assign indeterminate rather than false values to (1d) (and likewise to (1e)) might do so because its consequent is entailed by some but not all grounds, and this might be taken to violate a homogeneity presupposition. More generally, we think that deviations from the majority judgements need not be mistakes, but could have an interesting theoretical explanation.

Alternative explanations. One might hypothesize that (1d) is in principle equivalent to (1c) but that most participants misread "not both up" as "both not up" and therefore misinterpret (1d) as a paraphrase of (2). To control for this possibility, we tested the sentence *Switch A and switch B are not both up* in a pictorial context that shows switch A up and switch B down. 76.9% of speakers judged it true (N=290), suggesting that this particular source of noise is too weak to explain the observed effect. Another reason we reject this explanation is that it predicts that (1e) should be judged true, contrary to fact.

Alternatively, one might postulate silent material in the antecedents. An operator *Exh* (e.g. Fox, 2007; Spector, 2007) might strengthen the antecedent of (1c), for example by conjoining it with $\neg(\neg A \wedge \neg B)$. In effect, this operator would cause disjunction to be interpreted exclusively. This raises the question why the same effect is absent from (1d); and since it is absent, the conjunction problem still requires a solution. Furthermore, one would expect that this exhaustive strengthening applies to disjunctive main clauses at least

as often as it applies to disjunctive antecedents. This is not the case: In a separate survey (N=145), 81.4% of speakers judged the sentence *Switch A or switch B is down* true in a pictorial context that shows both switches down; only 15.9% judged it false, and 2.8% indeterminate. Finally, these results did not differ significantly ($p = 0.13$ on a χ^2 test) from how the sentence *Switch A and switch B are not both up* was judged in the same context (90.0% true, 8.5% false, 1.5% indeterminate, N=130). In other words, we have no evidence that exhaustive strengthening breaks de Morgan's law in main clauses. Therefore it is unlikely that exhaustive strengthening is the reason this law fails in disjunctive antecedents.

Conclusion. We have presented a set of data which challenge existing accounts of counterfactuals in two ways: first, they are incompatible with standard minimal change accounts of counterfactuals, regardless of the chosen similarity ordering; second, they show that counterfactuals whose antecedents are equivalent in classical logic can come apart in truth-conditions. Our account relies on two ideas: (i) disjunctive antecedents introduce multiple assumptions, while negative antecedents introduce only one; (ii) in making an assumption, one considers each way in which the assumption may obtain in the given scenario.

References

Alonso-Ovalle, L. (2009). Counterfactuals, correlatives, and disjunction. *Linguistics and Philosophy*, **32**(2), 207–244.

Ciardelli, I., Groenendijk, J., and Roelofsen (2013). Inquisitive semantics: A new notion of meaning. *Language and Linguistics Compass*, **7**(9), 459–476.

Erlewine, M. Y. and Kotek, H. (2016). A streamlined approach to online linguistic surveys. *Natural Language and Linguistic Theory*, **34**(2), 481–495.

Fine, K. (2012). Counterfactuals without possible worlds. *The Journal of Philosophy*, **109**(3), 221–246.

Fox, D. (2007). Free choice and the theory of scalar implicatures. In U. Sauerland and P. Stateva, editors, *Presupposition and Implicature in Compositional Semantics*, pages 71–120. Palgrave Macmillan, London, UK.

Kaufmann, S. (2013). Causal premise semantics. *Cognitive Science*, **37**(6), 1136–1170.

Lewis, D. (1973). *Counterfactuals*. Blackwell, Oxford, UK.

Lifschitz, V. (1990). Frames in the space of situations. *Artificial Intelligence*, **46**(3), 365–376.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press, Cambridge, UK.

Schulz, K. (2007). *Minimal models in semantics and pragmatics: Free choice, exhaustivity, and conditionals*. Ph.D. thesis, University of Amsterdam.

Spector, B. (2007). Aspects of the pragmatics of plural morphology: On higher-order implicatures. In U. Sauerland and P. Stateva, editors, *Presuppositions and implicature in compositional semantics*, pages 243–281. Palgrave, London, UK.