Antecedents of counterfactuals violate de Morgan's law

Lucas Champollion
champollion@nyu.edu
Joint work with Ivano Ciardelli and Linmin Zhang

LUSH talk (Utrecht)
November 13, 2015

1 Introduction

- Formulas (1a) and (1b) are equivalent in classical propositional logic (de Morgan's law), since they are true at the same possible worlds.

(1) a. $\neg A \vee \neg B$
    b. $\neg(A \wedge B)$

- In other systems, such as alternative semantics (Alonso-Ovalle, 2009) and inquisitive logic (Ciardelli, Groenendijk, and Roelofsen, 2013), they are not equivalent (see Figure 1).
- In inquisitive logic, an atomic proposition A denotes the set of all sets of A-worlds. (1a) denotes the set that contains:
  1. all sets of A-worlds and all sets of B-worlds
  2. but none of the remaining sets of $A \vee B$-worlds (such as the set of all worlds in which $A \vee B$ holds)
- As for (1b), it denotes a larger set: all the sets of $\neg(A \wedge B)$-worlds.
- That is, it contains all classical propositions that classically entail $(\neg A \vee \neg B)$ even if they don't tell us which disjunct is the case.
- In inquisitive logic, the denotations of the disjuncts can usually be recovered from the denotation of the disjunction. In propositional logic, they cannot.
- The starting point of our work is the question whether the second view (*recoverable disjuncts*) is needed.

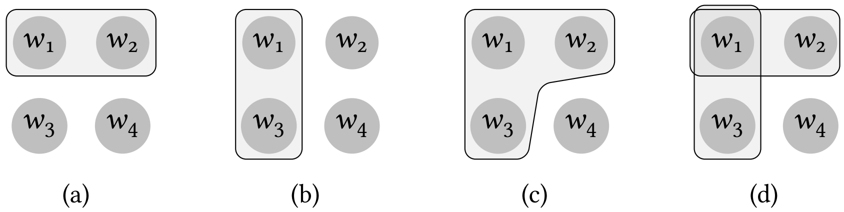

Figure 1: the disjunction of (a) and (b), as construed in truth-conditional semantics (c), and as construed in alternative and inquisitive semantics (d).

2 Counterfactuals

- Alonso-Ovalle (2009) suggested that the SDA law provide evidence for recoverable disjuncts

(2) Simplification of disjunctive antecedents (SDA)
  a. If Alice or Bob had come to the party, it would have been fun.
  b. Therefore, if Alice had come to the party, it would have been fun.

- The similarity-based account of counterfactuals by Lewis (1973) can't derive SDA unless it is patched up so the counterfactual can run separately on each disjunct.
- This essentially is the crux of Alonso-Ovalle's argument for recoverable disjuncts.
- But as we will see, these accounts are themselves problematic. We sought a stronger argument that does not rely on a particular account of counterfactuals.
- To this end, we decided to compare (1a) and (1b) directly in an experiment.

3 The switches scenario

- Schulz (2007):

  Imagine a long hallway with a light in the middle and with two switches, one at each end. One switch is called switch A and the other one is called switch B. As the following wiring diagram shows, the light is on whenever both switches are in the same position (both up or both down); otherwise, the light is off. (after Lifschitz, 1990)

- Test your intuitions (you may want to write them down):

(3)
  a. If switch A and switch B were both down, the light would be off. _____
  b. If switch A and switch B were not both up, the light would be off. _____
  c. If switch A was down, the light would be off. _____
  d. If switch A or switch B was down, the light would be off. _____

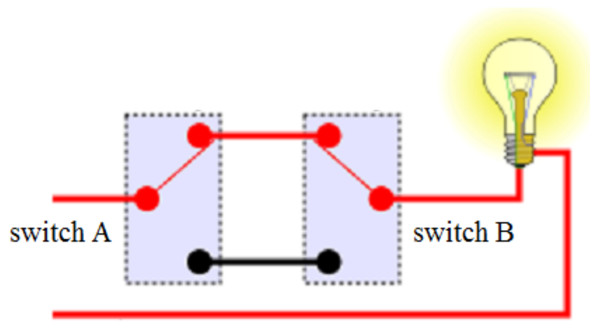

Figure 2: Lifschitz' switches

4 Our experiment

- We used MTurk to collect truth value judgments in the context shown above (including the picture).
- Participants could choose between *true*, *false* and *indeterminate*.
- We were mainly interested in these two sentences:

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

- We intuitively felt a difference between these sentences: we were much more hesitant to judge (4a) true than (4b).
- To control for naturalness, we ran a pre-test and found that the two sentences did not significantly differ in naturalness as judged on a 7-point scale with no scenario supplied (N=55).
- We also tested the following sentences:

(5) a. If switch A was down, the light would be off. $(\neg A) > \text{OFF}$
b. If switch B was down, the light would be off. $(\neg B) > \text{OFF}$

- From our own intuitions, we expected that these sentences should both be judged true.

Finally we also tested this sentence:

(6) a. If switch A and switch B were not both up, the light would be on. $\neg(A \wedge B) > \text{ON}$

- We expected this to be judged false or indeterminate.
- We only showed people one of the sentences above at a time.
- We also showed everyone a filler sentence that we take to be uncontroversially false:

(7) If switch A and switch B were both down, the light would be off. $(\neg A \wedge \neg B) > \textsc{off}$

- All our subjects were based in the US. We eliminated subjects who participated more than once or didn't finish (less than 1%), didn't speak American English natively (about 4%), or didn't judge the filler sentence false (about 38%).
- The results are shown in Table 1.

Table 1: Results

Conditions	Total number	True	(%)	False	(%)	Indet.	(%)
$(\neg A) > \textsc{off}$	255	**169**	**66.27%**	6	2.35%	80	31.37%
$(\neg B) > \textsc{off}$	234	**153**	**65.38%**	7	2.99%	74	31.62%
$(\neg A \vee \neg B) > \textsc{off}$	346	**242**	**69.9%**	12	3.5%	92	26.6%
$\neg(A \wedge B) > \textsc{off}$	356	80	22.5%	129	36.2%	**147**	**41.3%**
$\neg(A \wedge B) > \textsc{on}$	200	43	21.5%	63	31.5%	**94**	**47.0%**

- $(\neg A) > \textsc{off}$, $(\neg B) > \textsc{off}$, and $(\neg A \vee \neg B) > \textsc{off}$ were all judged true. By contrast, $\neg(A \wedge B) > \textsc{off}$ was generally judged false or indeterminate. This difference was highly significant ($p < 0.0001$ on a chi-square test).
- The similarity between $(\neg A) > \textsc{off}$, $(\neg B) > \textsc{off}$, and $(\neg A \vee \neg B) > \textsc{off}$ is striking. None of these conditions differed significantly.

5 Discussion

- We draw the following pretheoretical conclusions from the experiment:
- "If A or B then C" is interpreted in the same way as "If A then C, and if B then C", in accordance with SDA.
- When "If A or B then C" is interpreted, A and B are changed only one at a time. We consider two counterfactual scenarios: what if A but not B; what if B but not A.
- When "If (not both A and B) then C" is interpreted, we consider three counterfactual scenarios: what if A but not B; what if B but not A; what if neither A nor B.
- These results constrain our theories: De Morgan's law does not hold in the antecedents of counterfactuals.
- Propositional logic needs to be either supplemented or replaced in order to account for this.

6 Quick primer on counterfactuals

(8) If kangaroos had no tails, they would topple over. (Lewis, 1973)

- Material implication at the actual world is not an option.
- Nor is material implication at all possible worlds:

(9) a. If kangaroos had no tails but used crutches, they would topple over.
b. $A > C \not\Rightarrow (A \wedge B) > C$

  - There will generally be some A-worlds which are very odd and remote
  - It should not matter whether such worlds are also C-worlds

- Stalnaker (1968), Lewis (1973): worlds are ordered based on how similar they are to the actual world
- "In any possible state of affairs in which kangaroos have no tails, and which resembles our actual state of affairs as much as kangaroos having no tails permits it to, the kangaroos topple over." (Lewis, 1973)
- Simplifying Lewis's proposal a bit, it says:
  - $A > C$ is true iff C is true at every closest A-world.
- In case there is exactly one closest A-world, this amounts to:
  - $A > C$ is true iff C is true at the closest A-world (Stalnaker, 1968)
- One problem: counterfactuals involving drastic changes should never be true (Fine, 1975)

(10) a. If Nixon had pressed the button, there would have been a nuclear holocaust.
b. The closest A-worlds will be worlds where the wire is cut etc.

- For this to work, a cut wire must mean a bigger difference than a nuclear explosion.
- Another problem (Schulz, 2007): similarity needs to conspire to simulate causal effects

(11) a. If switch A was down, the light would be off. *true*
b. If switch A was down, switch B would be down as well. *false*

- For this to work, a changed switch must mean a bigger difference than a changed light.
- Only the light causally depends on switch A. Switch B does not. But this fact is not relevant for Stalnaker and Lewis.
- Lewis (1979) responded by introducing a system of weights that makes the nuclear-holocaust world more similar to ours than the world in which the wire is cut.

- But our results are incompatible with the Stalnaker/Lewis system on any similarity metric.

(12) a. Most of our speakers judged $(\neg A) > C$ and $(\neg B) > C$ true.
b. So the closest $\neg A$-worlds and the closest $\neg B$-worlds are C-worlds.
c. Most of our speakers judged $(\neg(A \wedge B)) > C$ false or indeterminate.
d. So not every closest $\neg(A \wedge B)$-worlds can be a C-world.
e. But every closest $\neg(A \wedge B)$-world is either a closest $\neg A$-world or a closest $\neg B$-world, and by (12b) must be a C-world.

- Pearl (2000), Schulz (2007), Kaufmann (2013), and Briggs (2012) replace the similarity ordering on worlds by a causal network (a directed acyclic graph over propositions).

- I will describe Kaufmann's system here because it builds on Kratzer's familiar premise semantics, and because it can be extended to account for our results.

7 Simplified version of Kaufmann (2013)

- Causal dependencies are described by a causal structure and a set of causal laws.

- The causal structure is a directed acyclic graph whose nodes (the *variables*) are partitions over the set of possible worlds.

- Figure 7 shows the causal structure for the switches scenario. Its variables are "whether switch A is up" (independent), "whether switch B is up" (independent), and "whether the light is on" (dependent).

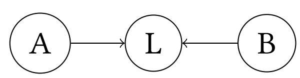

Figure 3: The causal structure for the switches scenario

- The cells of a variable are called *settings*: e.g. a ("switch A is up"), $\overline{a}$ ("switch A is down").

- A set of variable settings that are true in the actual world is called a *premise set*. A *causal premise set* is a premise set that is closed under the ancestor relation.

- To evaluate a counterfactual, we first compute the *causal premise background*: the set of all causal premise sets.

- Suppose both switches are up and the light is on. Our true variable settings are a, b, and l. Our causal premise background is $\{\emptyset, \{a\}, \{b\}, \{a, b\}, \{a, b, l\}\}$.

- Then we identify the set of all those causal premise sets that are consistent with the antecedent.

- If our antecedent is "if switch A was down" ($\overline{a}$), this set is $\{\emptyset, \{b\}\}$.
- When considering an antecedent we typically hold constant all facts that don't causally depend on it:

(13) If switch A was down, switch B would still be up. *true*

- Accordingly, we are only interested in the *maximal* causal premise sets. Of $\emptyset$ and $\{b\}$ only $\{b\}$ is maximal.
- We call the set of maximal causal premise sets a *modal base*.
- In the simplest case, causal laws are material conditionals whose arrows go from parents to children in the causal network:

(14) a. $(a \wedge b) \rightarrow l$
b. $(\overline{a} \wedge b) \rightarrow \overline{l}$
c. $(a \wedge \overline{b}) \rightarrow \overline{l}$
d. $(\overline{a} \wedge \overline{b}) \rightarrow l$

- $A > C$ is true iff for every set S in its modal base,[1] C is entailed by the conjunction of A, the propositions in S, and the causal laws that S is consistent with.
- We need to conjoin our antecedent $\overline{a}$ with our sole maximal set, $\{b\}$, and with the four laws, in particular $(\overline{a} \wedge b) \rightarrow \overline{l}$. The result is equivalent to $\overline{a} \wedge b \wedge \overline{l}$.
- The sentence *If switch A was down, the light would be off* is predicted true.

8 Extending Kaufmann (2013) to complex antecedents

- To model our results we will need to modify the procedure somewhat.
- We have seen that we got the same results for $\neg a \vee \neg b$ as we did for $\neg a$ and for $\neg b$.
- This suggests that a separate instance of the counterfactual is run on each disjunct.
- Suppose we have a counterfactual operator $A \blacktriangleright C$ that is defined only for noninquisitive meaning, e.g. that of Stalnaker, Lewis, or Kaufmann.
- We can lift it into inquisitive semantics in the following way (where $\textsc{Alt}(\phi)$ returns the set of maximal subsets of ϕ):

(15) Inquisitive counterfactual conditional
$[\![\phi > \psi]\!] = \{p \,|\, \text{for all } A \in \textsc{Alt}(\phi) \text{ there is a } C \in \textsc{Alt}(\psi) \text{ such that } p \subseteq A \blacktriangleright C\}$

[1] Kaufmann only considers those sets which are compatible with as many laws as possible (*ordering source*). We can ignore this step in our scenario because no causal laws ever get broken.

- In the cases of interest, ψ has only one alternative C ("the light is off"), but ϕ may have two alternatives (A_1 ="switch A is down", A_2 = "switch B is down").
- Then this definition amounts to separately testing $A_i \blacktriangleright C$ for each A_i.
- The case of $\neg(a \wedge b)$ needs more work: we have to change Kaufmann's operator itself.

(16) If switch A and switch B were not both up, the light would be off. – *judged false*

- As before, our causal premise background is $\{\emptyset, \{a\}, \{b\}, \{a, b\}, \{a, b, l\}\}$.
- Given this, the modal base of (16) is $\{\emptyset, \{a\}, \{b\}\}$, of which $\{a\}$ and $\{b\}$ are maximal.
- Kaufmann would now conjoin each of these sets with the antecedent and the causal laws:

(17) a. $a \wedge \neg(a \wedge b) \wedge [laws] \equiv a \wedge \overline{b} \wedge \overline{l}$
b. $b \wedge \neg(a \wedge b) \wedge [laws] \equiv \overline{a} \wedge b \wedge \overline{l}$

- In each of these cases, the light is off. So the sentence is predicted true, contrary to fact.
- When we built the modal base, we treated the antecedent A as a whole.
- Instead we will consider separately each of the three "ways of making A true", or what we will call the grounds for A: $\{a, \overline{b}\}, \{\overline{a}, b\}, \{\overline{a}, \overline{b}\}$. This forces attention to the case where both switches are down.
  - Given a proposition ϕ, a set V of partitions over worlds (i.e. of causal variables), and a nonempty set S of settings of some of these variables, S *controls* ϕ iff $\bigcap S$ entails ϕ but there is a set S' of settings of the same variables such that $\bigcap S'$ is inconsistent with ϕ. *S is a ground for ϕ* (a "way of making ϕ true") if S is minimal among the sets that control ϕ.
  - Let $\phi = [\![a]\!]$, $V = \{A, B, L\}$, $S = \{a, b\}$, $S' = \{\overline{a}, b\}$. Now $\bigcap S$ entails ϕ but $\bigcap S'$ is inconsistent with ϕ. Therefore S controls ϕ. But so does $\{a\}$, hence S is not a ground for ϕ. Rather, $\{a\}$ is the only ground for ϕ.
  - Let $\phi = [\![\neg(a \wedge b)]\!]$, $V = \{A, B, L\}$, $S_1 = \{a, \overline{b}\}$, $S_2 = \{\overline{a}, b\}$, $S_3 = \{\overline{a}, \overline{b}\}$, $S' = \{a, b\}$. Now each $\bigcap S_i$ entails ϕ but $\bigcap S'$ is inconsistent with ϕ. Therefore each S_i controls ϕ. As for $\{a\}, \{b\}, \{\overline{a}\}, \{\overline{b}\}$, none of them controls ϕ, hence each S_i is a ground for ϕ.
- For each ground of A, we separately compute the set of maximal causal premise sets consistent with it.

(18) a. For $\{a, \overline{b}\}$, that set is $\{\{a\}\}$.
b. For $\{\overline{a}, b\}$, that set is $\{\{b\}\}$.
c. For $\{\overline{a}, \overline{b}\}$, that set is $\{\emptyset\}$.

- Our updated recipe is now as follows:

- $A > C$ is true iff for each ground G of A, for each maximal causal premise set M consistent with G, the causal laws together with G and M entail C.
- Now (16) requires us to consider three grounds for the antecedent:

(19) a. Ground 1: Switch A is up and switch B is down.
b. Its only maximal causal premise set says that switch A is up.
c. Given the causal laws, the light is off.

(20) a. Ground 2: Switch A is down and switch B is up.
b. Its only maximal causal premise set says that switch B is up.
c. Given the causal laws, the light is off.

(21) a. Ground 3: Both switches are down.
b. Its only maximal causal premise set does not say anything.
c. Given the causal laws, the light is on.

- Since on ground 3 the light is on, (16) is predicted false, as desired.

9 What about implicatures?

- We now have an account of our results, and it relies on recoverable disjuncts. But it is certainly not the only possible account.
- What if *or* means XOR?

(22) If switch A XOR switch B was down, the light would be off.
$\Leftrightarrow$ If switch A or switch B was down but not both, the light would be off.

- This would explain why *not A or not B* is not equivalent to *not (A and B).*
- Now, *or* certainly does not *literally* mean XOR:

(23) It is not the case that switch A XOR switch B is down.
$\Leftrightarrow$ Either they are both up or they are both down.

- But what if *or* is enriched from $\vee$ to XOR by implicature? Cf. the top-level implicature in a disjunction:

(24) Switch A or switch B is down.
$\rightsquigarrow$ They are not both down.

- Maybe the antecedent is locally strengthened to *only A or B.*

(25) $(\neg A \vee \neg B) > C$ is interpreted as $(\text{EXH}(\neg A \vee \neg B)) > C$, where EXH $\approx$ "only".

- Presumably, what breaks the symmetry is this inequality:

(26) $\text{EXH}(\neg A \vee \neg B) \neq \text{EXH}(\neg(A \wedge B))$

- This predicts that a certain implicature should *not* hold:

(27) Switch A and switch B are not both up.
$\not\rightsquigarrow$ They are not both down.

- We are not sure that this implicature is absent. But suppose it is, (26) shows that in a sense, de Morgan's law fails to hold even at the root level – but it only shows up as implicatures.
- There is a question how to derive (26). One possibility is that EXH has no effect on $\neg(A \wedge B)$.
- This depends on what the syntactic alternatives of that expression are.

10 Conclusion

- Our experiment shows that de Morgan's law fails to hold in the antecedents of counterfactuals.
- It also fails in inquisitive semantics. This provides us with an intuitive explanation for the contrast, which is then put to work in our semantics.
- Our experiment also shows that $(\neg A) > C$ and $(\neg B) > C$ do not entail $(\neg(A \wedge B)) > C$, contra the Stalnaker/Lewis account.
- An antecedent of the form $\neg(A \wedge B)$ invites contemplation of three "grounds": A but not B, B but not A, and crucially, neither A nor B.
- An antecedent of the form $\neg A \vee \neg B$ only invites contemplation of the first two cases.
- We have also provided empirical support for SDA: $(A \vee B) > C$ entails $A > C$. This is a good fit for inquisitive semantics and other frameworks with recoverable disjuncts.

References

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

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

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

Fine, Kit (1975). Critical Notice. *Mind* 84(335), 451–458. DOI: 10.1093/mind/LXXXIV.1.451.

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

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

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

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

Pearl, Judea (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/cbo9780511803161.

Schulz, Katrin (2007). Minimal models in semantics and pragmatics: Free choice, exhaustivity, and conditionals. PhD thesis. Universiteit van Amsterdam.

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