Two switches in the theory of counterfactuals

A study of truth conditionality and minimal change

Lucas Champollion (NYU) · champollion@nyu.edu
Joint work with Ivano Ciardelli (UvA) and Linmin Zhang (Concordia)

Workshop on Logic and Algorithms in Computational Linguistics 2017
Department of Mathematics, Stockholm University · 8/16-8/19, 2017

Manuscript available at http://ling.auf.net/lingbuzz/003200

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

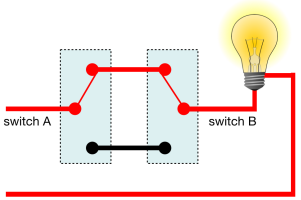


Figure 1: Switch A and B are both up, and the light is on.

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

(2) a. If switch A was down, the light would be off. T F I
b. If switch B was down, the light would be off. T F I
c. If switch A or switch B was down, the light would be off. T F I
d. If switch A and switch B were not both up, the light would be off. T F I
e. If switch A and switch B were not both up, the light would be on. T F I

(2a) = $\overline{A} > \textsc{off}$; (2b) = $\overline{B} > \textsc{off}$; (2c) = $\overline{A} \vee \overline{B} > \textsc{off}$; (2d) = $\neg(A \wedge B) > \textsc{off}$; (2e) = $\neg(A \wedge B) > \textsc{on}$

Table 1: Results of the main experiment

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

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

Table 2: Results of pretest (picture and text: both switches are down)

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 3: Changed picture and text: the light is on only if the switches are up

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

- Do the truth conditions of a sentential clause completely determine its meaning?
  ⇝ **No**; otherwise $\overline{A} \vee \overline{B} > \textsc{off}$ and $\neg(A \wedge B) > \textsc{off}$ should be equivalent
- Does the interpretation of counterfactuals with complex antecedents conform to the minimal change requirement?
  ⇝ **No**; otherwise $\overline{A} > \textsc{off}$ and $\overline{B} > \textsc{off}$ should jointly entail $\neg(A \wedge B) > \textsc{off}$

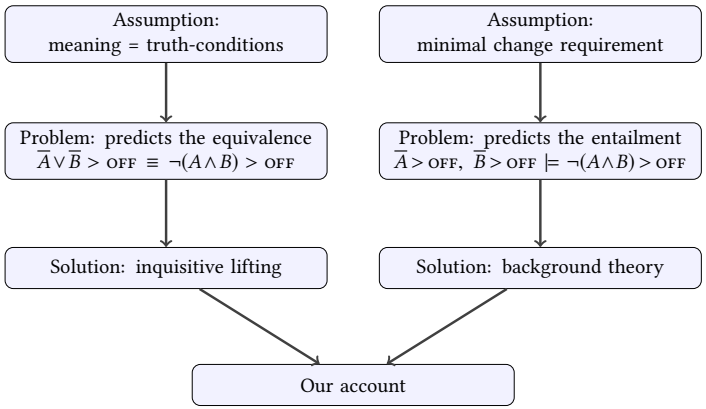
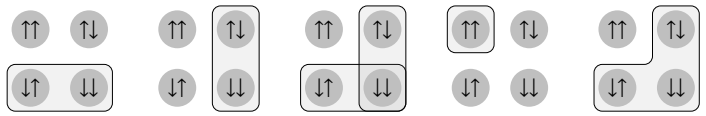


Figure 2: The theory at a glance.



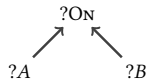
(a) $\overline{A}$ (b) $\overline{B}$ (c) $\overline{A} \vee \overline{B}$ (d) $A \wedge B$ (e) $\neg[A \wedge B]$

Figure 3: Inquisitive alternatives of some simple sentences. ↑↑ represents a world where both switches are up, ↑↓ a world where A is up but B is down, etc.

Definition 1 (Inquisitive lifting of a base theory "$\Rrightarrow$" of counterfactuals)**.**
$s \models \varphi > \psi$ iff $\forall p \in \text{Alternatives}(\varphi)\ \exists q \in \text{Alternatives}(\psi)$ such that $s \subseteq (p \Rrightarrow q)$

Definition 2. (Causal model) A causal model is a pair $M = \langle V, L \rangle$ consisting of:

- A set V of *causal variables* (partitions over the set of worlds). We call the partition cells *settings*. The *value* of a variable X at w, denoted X_w, is its true setting at w. These notions are generalized to sets of variables in the obvious way.
- A set L of *laws*. A law is a tuple $\langle C_l, E_l, m_l \rangle$ where C_l, the *cause set*, is a set of causal variables; E_l, the *effect*, is another causal variable; and m_l, the *map*, is a partial function from settings of C_l to settings of E_l. The *upshot of* l, written $|l|$, is the set of worlds at which the causal law in question is obeyed. The laws induce a *causal graph* that connects the causal variables.



Given a causal model M and a world w, we define a base theory of counterfactuals as follows:

Definition 3 (Facts)**.**
A fact is a setting of a causal variable which is true at w.

Definition 4 (Facts that contribute to the falsity of a proposition)**.**
A fact f contributes to the falsity of a proposition a in case some set of facts is consistent with a but not with $a \cap f$ (or equivalently, if there are finitely many facts: there is a maximal set of facts that is consistent with a and that does not contain f).

Definition 5 (Factual background map)**.**
A factual background map is a function $\mathcal{B}$ that maps any proposition a and world w to a set $\mathcal{B}(w, a)$ of facts at w such that neither f itself nor any ancestor of f in the causal graph of M contributes to the falsity of a.

Definition 6 (Intervention)**.**
A proposition a intervenes on a law l at a world w in case the value of the effect of l contributes to the falsity of a at w.

Definition 7 (Law background for a proposition)**.**
The law background for a at w, denoted $\mathcal{L}(w, a)$, is the set of all upshots $|l|$ of laws $l \in L$ on which a does not intervene at w.

Definition 8 (Hypothetical context created by an assumption)**.**
Let $\mathcal{B}$ be a factual background map. The hypothetical context given by an assumption a at world w under $\mathcal{B}$, denoted $f_{\mathcal{B}}(w, a)$, is the intersection of all propositions in $a \cup \mathcal{B}(w, a) \cup \mathcal{L}(w, a)$.

Definition 9 (Truth conditions for counterfactuals)**.**
A conditional proposition $a \Rrightarrow c$ is true under a factual background map $\mathcal{B}$ just in case $f_{\mathcal{B}}(w, a) \subseteq c$.

Acknowledgments

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