# Explaining negative counterfactuals

Muyi Yang

Conditionals workshop @ UConn, April 6, 2019

- Languages like Mandarin has negative counterfactuals (NCs) that are headed by the complementizer *yaobushi* (lit. if-NEG-COP) 'if not'.

(1)

| yaobushi | xia | yu, | di | shang | jiu | bu | hui | shi | le. |
|---|---|---|---|---|---|---|---|---|---|
| IF.NOT | fall | rain | ground | on | then | NEG | will | get.wet | ASP |

'If it had not rained, the streets would not have been wet.'

- In Ippolito and Su (2014) and Yang (ta), it is observed that Mandarin NCs (i) reject indicative interpretation, (ii) reject backtracking interpretations, (iii) and presuppose the truth of the prejacent of *yaobushi* as well as the falsity of the consequent.

- Note that standard positive counterfactuals with sentential negation in antecedents do not have these properties.

(2)

| yaoshi | mei | xia | yu, | di | shang | jiu | bu | hui | shi | le. |
|---|---|---|---|---|---|---|---|---|---|---|
| IF | NEG.PERF | fall | rain | ground | on | then | NEG | will | get.wet | ASP |

'If it had not rained, the streets would not have been wet.'

- Apart from Mandarin, the construction of NCs seems quite robust crosslinguistically, i.e. English *if not for*, *if it weren't for*, Spanish *Si no fuera porque* (lit. 'if not were because') (Henderson 2010, 2011), and Tagalog *kundi* 'if not that' (Nevins 2002).

(3) If not for the jacket he'd thought to put on before leaving home, he'd be drenched from this deluge of rain. (COCA)

(4) If it weren't for the photo album, I wouldn't even remember their faces. (COCA)

- No significant contrast between my intuition about Mandarin NCs and English speakers' intuition about English NCs (but feel free to disagree and/or think about NCs in other languages), English data only today.

# 1 Negative counterfactuals in discourse

The intuition that I want to explore

While standard counterfactuals can be used to address various issues in discourse, negative counterfactuals (henceforth NCs) are only used for explaining facts.
Specifically, I hope to tease apart two discourse functions: EXPLANATION OF FACTS (henceforth EXPLANATION) and COUNTERFACTUAL DESCRIPTION (henceforth DESCRIPTION).

(5) ***Assassin's poison:*** *Assassin poisons Victim's coffee. Victim drinks it and dies. But if Assassin hadn't poisoned the coffee, Backup would have, and Victim would have died anyway.* (Modified from Hitchcock's 2007 'Early Preemption')

a. Q: How come Victim died?
A: If there had been no poison in the coffee, Victim would not have died. (EXPLANATION)

b. Q:What would have happened if there had been no poison in the coffee?
A: If there had been no poison in the coffee, Victim would not have died. (DESCRIPTION)

- A standard counterfactual '*if A, would C*' can address EXPLANATION: for the question 'Why was it not the case that ¬*C* occurred?', we learn that the reason is ¬*A*.

- A standard counterfactual '*if A, would C*' can also address DESCRIPTION: for the question 'What would have happen if *A* had been the case?', we learn that *C* would have occurred in that counter-to-fact hypothetical situation.

- Not all counterfactuals serve these purposes, e.g. semifactual counterfactuals:

(6) ***Assassin's poison***

a. Q: How come Victim died?
A: # If Assassin hadn't poisoned the coffee, Victim would still have died. (#EXPLANATION)

b. Q: What would have happened if Assassin hadn't poisoned the coffee?
A: If Assassin hadn't poisoned the coffee, Victim would still have died.(✓DESCRIPTION)

- **My overarching claim: NCs '*if not for p, would q*' only address EXPLANATION.**

(7) ***Assassin's poison:***

a. Q: How come the Victim died?
A: If not for there being poison in the coffee, Victim would not have died. (✓EXPLANATION)

b. Q: What would have happened if there had been no poison in the coffee?
A: #If not for there being poison in the coffee, Victim would not have died. (#DESCRIPTION)

(8) #If not for Assassin having poisoned the coffee, Victim would still have died.

- To make this intuition more concrete, I will show two specific cases where standard and negative counterfactuals come apart. The two cases demonstrate:
  - when the need for an EXPLANATION could arise; and

  - what specific types of EXPLANATION an NC permits.

  I capture these observations by enriching a causal model with a normalcy ordering (see the philosophy literature on the connection between explanations and expectations such as Gärdenfors 1988).

- I assume that EXPLANATION builds on the notion of causality (Pearl 2000, Halpern 2016 a.o.). The two previous works on NCs are not entirely satisfactory in this respect:
  - Ippolito and Su (2014) offer a syntax-semantic analysis based on the assumption that a type-mismatch drives movement of negation into the complementizer, not addressing the causality-based constraints on NCs;
  - Henderson (2010, 2011) recognizes the importance of causality in the semantics of NCs and provides an analysis based on Schulz's (2007) causal model. But it will become clear that compared with an EXPLANATION-based analysis, a mere causality-based analysis overgenerates the felicity of NCs in certain contexts.

## 2 When does EXPLANATION arise?

(9) ***Anaerobic Room:*** *The air of the room is vacuumed and now there is no oxygen. The match is struck, but does not light.*

a. If there had been some oxygen, the match would have lit.

b. If not for there having been no oxygen, the match would have lit.

(10) ***Active Aerobic Room:*** *We enter a normal room where oxygen is present. John comes in and strikes the match, and the match lights.*

a. If the match had not been struck, it would not have lit.

b. If not for the match having been struck, it would not have lit.

(11) ***Inactive Aerobic Room:*** *We enter a normal room where oxygen is present. But no one strikes the match, and nothing happens.*

a. If the match had been struck, it would have lit.

b. #If not for the match not having been struck, it would have lit.

- Compare (11b) in the ***Inactive Aerobic Room*** vs. (12b) the ***Match-striker's Aerobic Room***.

(12) ***Match-striker's Aerobic Room:*** *We enter a normal room where oxygen is present, and there is always a match striker whose job is to strike any match if there is any. Strangely, the match striker is absent today. The match does not light.*

a. If the match had been struck, it would have lit.

b. If not for the match not having been struck, it would have lit.

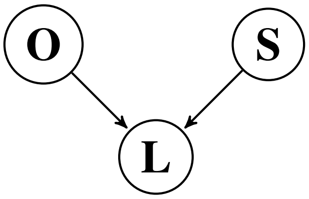
- Core intuition: The utterer of an NC needs to know that *something unexpected* has occurred in order to seek for an explanation. In the ***Inactive Aerobic Room***, nothing happens and nothing needs to be explained, unless we change what our expectations are, as in the

***Match-striker's Aerobic Room***.
Note: In ***Inactive Aerobic Room***, NC is degraded although the intended reading is an ontic one that follows the flow of causality, i.e. Henderson (2010, 2011) will make incorrect predictions.

## 2.1 Ingredients and first attempt of implementation

Let a context $c$ be a quadruple $\langle w, \mathscr{C}, f, g\rangle$.

- $w$ is the world of $c$.
- $\mathscr{C}$ is the salient causal structure in the form of $\langle U, <\rangle$ (Kaufmann 2013). $U$ is a set of partitions on $W$, and $<$ is a directed acylic graph over U, e.g. $P_1 < P_2$ indicates that there is a non-empty causal path from $P_1$ to $P_2$. For instance, (13a) is the DAG I assume for the match scenarios. $O$, $S$ and $L$ are bipartitions on $W$ that can be paraphrased by questions 'Is there oxygen?' 'Is the match struck?' and 'Does the match light?'.
- $f$ is the Kratzer-style epistemic modal base. For simplicity, I assume that $f(w)$ contains only the causally-relevant truths (those denoted by the cells in the partitions in $U$) known by the agent, and that the agent knows the answers for the variables in $U$ (i.e. the causal equations that can be paraphrased by the biconditional in (13b)).

(13) a. 

b. $s \wedge o \leftrightarrow l$

- $g$ is a Krater-style ordering source that orders worlds according to what the speaker takes to be normal (or the 'expectation pattern' of Veltman 1996): $u \leq_{g(w)} v$ iff the world $u$ is at least as expected as $v$ in terms of the norms determined by $g(w)$.

(14) a. '*if not for p, would q*' is defined in $c$ only if the cell of $w$ is not the best possibility (in the sense of Kratzer 1981's comparative possibility[1]) among the cells in the partitions in $U$.

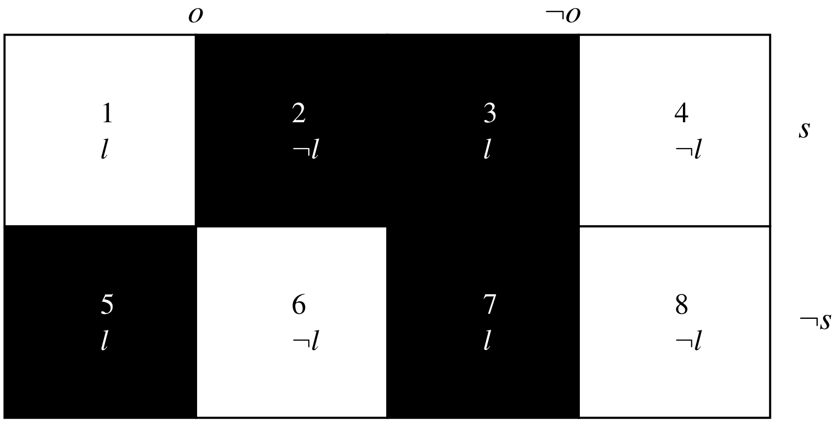
b. Once defined, '*if not for p, would q*' is true in $c$ iff $q$ follows from a counterfactual revision of the agent's belief with $\neg p$ that is sensitive to $\mathscr{C}$ and $f$. That is, once the agent's belief is updated with $\neg p$, $p$ and its causal descendants are removed, and $q$ follows.

[1] The definition of comparative possibility from Kratzer (1981): A proposition $p$ is at least as good a possibility as a proposition $q$ in a world $w$ with respect to a modal base $f$ and an ordering source $g$ iff for all $u$ such that $u \in \cap f(w)$ and $u \in q$, there is a $v \in \cap f(w)$ such that $v \leq_{g(w)} u$ and $v \in p$.

## 2.2 Two toy models

For ***<u>Anaerobic Room</u>*** (9), ***<u>Active Aerobic Room</u>*** (10) and ***<u>Inactive Aerobic Room</u>*** (11).

(15) The partition $U$ on $W$ (I refer to the number of each cell as the set of worlds that are in that cell, e.g. the 1-worlds):

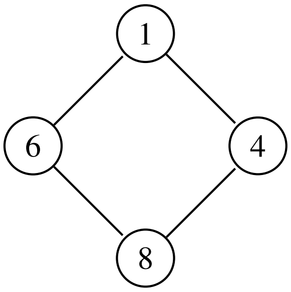


| $o$ | | $\neg o$ | | |
|---|---|---|---|---|
| 1<br>$l$ | ■ 2<br>$\neg l$ | ■ 3<br>$l$ | 4<br>$\neg l$ | $s$ |
| ■ 5<br>$l$ | 6<br>$\neg l$ | ■ 7<br>$l$ | 8<br>$\neg l$ | $\neg s$ |

The cells in black are not in the modal base, because they violate the causal biconditional (13b).

(16) $f_{\textit{Anaerobic}}(w) = \{\neg o, s, \neg l\}$, the world of $c_{\textit{Anaerobic}}$ is thus a 4-world;
$f_{\textit{Inactive-Aerobic}}(w) = \{o, \neg s, \neg l\}$, the world of $c_{\textit{Inactive-Aerobic}}$ is thus a 6-world;
$f_{\textit{Active-Aerobic}}(w) = \{o, s, l\}$, the world of $c_{\textit{Active-Aerobic}}$ is thus a 1-world.

(17) In these scenarios, I assume that the speaker takes it normal that there is oxygen, and that there is no match striker. Thus the sets of worlds of the cells can be ranked in terms of comparative possibility, visualized by a rightward path, e.g. the set of 6-worlds is strictly better a possibility than any other set of cell-worlds, the set of 1-worlds is strictly better a possibility than the set of 4-worlds.



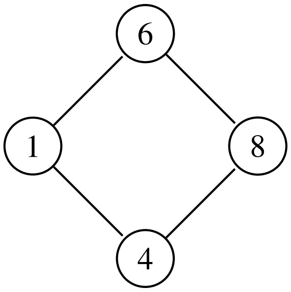
**(11b) is infelicitous because it violates the presupposition (14b): the world of $c_{\textit{Inactive-Aerobic}}$ is too optimal with respect to the normalcy ordering. No unexpectedness, thus no need for an explanation.**

For ***<u>Match striker's Aerobic Room</u>*** (12).

(18) The partition $U$ on $W$: same as (15)

(19) $f_{\textit{Match-striker's-Aerobic}}(w) = \{o, \neg s, \neg l\}$, the world of $c_{\textit{Match-striker's-Aerobic}}$ is thus a 6-world (same as the world of $c_{\textit{Inactive-Aerobic}}$).

(20) In this scenario, I assume that the speaker takes it normal that there is oxygen, and that there is a match striker. The ordering of cells:



**(12b) is felicitous because the world of $c_{\textit{Match-striker's-Aerobic}}$ is not too as expected.**

## 3 Which EXPLANATION to choose?

(21) ***Forgetful Bodyguard***: *Bodyguard possesses an antidote that neutralizes any type of poison, but does no harm if taken alone. She administers the antidote every day to whatever Victim eats or drinks. One day, Assassin poisons Victim's coffee, but Bodyguard happens to forget to administer antidote. Victim dies.* (Modified from 'Omission' in Hitchcock 2007)

a. If Bodyguard hadn't forgot to administer the antidote, Victim wouldn't have died.

b. If not for Bodyguard having forgot to administer the antidote, Victim wouldn't have died.

(22) a. If Assassin hadn't poisoned the coffee, Victim wouldn't have died.

b. ?If not for Assassin having poisoned the coffee, Victim wouldn't have died.

- Core intuition: In a collider model where one variable causally depends on multiple causal ancestors (Pearl 2000), one fact can be explained by multiple causes. The contrast between (21) and (22) shows that an NC tends to pick out the *normalcy-wise good* proposition as its antecedent.
  Note again: In (22), NC is degraded although the intended reading is an ontic one that follows the flow of causality. Again, Henderson (2010, 2011) would make incorrect predictions.

- This idea has been advocated by philosophers like Sintonen (1984), who argued that explanans play the role of reducing the value of 'surprise' by updating the speaker's knowledge.

(23) An EXPLANATION-based semantics of NCs: Let $c = \langle s, \mathscr{C}, \varepsilon \rangle$

a. '*if not for p, would q*' is defined in $c$ only if, the cell of $w$ is not the best possibility (in the sense of Kratzer 1981's comparative possibility) among the cells in the partitions in $U$.

b. $p$ is such that updating $f(w)$ with $\neg p$[2] leads to a cell that is strictly a better possibility than the cell that $w$ is in.

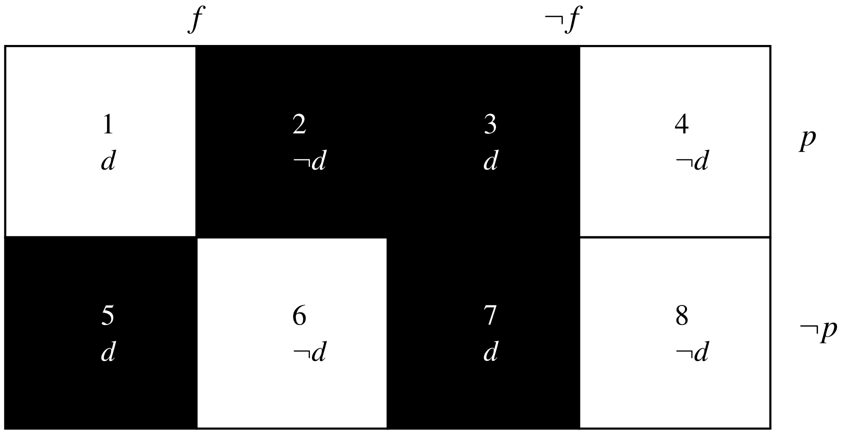
[2] In the sense of Kratzer-style premise semantics, i.e. $f[\neg p](w) = f'(w) \cup \{\neg p\}$ where $f'(w)\cup$ is the maximal subset of $f(w)$ that is logically consistent with $\neg p$.

c. Once defined, '*if not for p, would q*' is true in $c$ iff $q$ follows from a counterfactual revision of the agent's belief with $\neg p$ that is sensitive to $\mathscr{C}$ and $f$. That is, once the agent's belief is updated with $\neg p$, $p$ and its causal descendants are removed, and $q$ follows.

## 3.1 Toy model

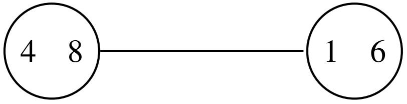
For ***Forgetful Bodyguard*** (21) and (22).

(24) The partition $U$ on $W$:



(25) $f_{\textit{\textbf{Forgetful-Bodyguard}}}(w) = \{f, p, d\}$, the world of $c_{\textit{\textbf{Forgetful-Bodyguard}}}$ is thus a 1-world.

(26) In this scenario, I assume that the speaker takes it normal that the Bodyguard does not forget to administer that antidote.[3] Thus the ranking of the cells:



**(22b) is degraded because the presupposition (23b) fails:**

- $f_{\textit{\textbf{Forgetful-Bodyguard}}}[\neg p](w) = \{\neg p, f\}$,
- After this update, the closest world to the world of $c_{\textit{\textbf{Forgetful-Bodyguard}}}$ is a 6-world. The 6-world cell is not strictly as better a possibility than the 1-world cell is.

**In contrast, (21b) is a perfect NC:**

- $f_{\textit{\textbf{Forgetful-Bodyguard}}}[\neg f](w) = \{\neg f, p\}$,
- After this update, the closest world to the world of $c_{\textit{\textbf{Forgetful-Bodyguard}}}$ is a 4-world. The 4-world cell is strictly as better a possibility than the 1-world cell is.

[3] In the two models for the match-examples, both the oxygen and the match-striker are normalcy-relevant. For now, I do not have a good motivation for not assuming that the speaker takes it normal that Assassin does or does not poison. In fact, this assumption would yield the wrong result.

# 4 Summary, predictions and some more puzzles

What's new in the implementation:

- Compare with the model of graded causation in Halpern (2016)'s §3, I took into account knowledge, which is a crucial connection in capturing EXPLANATION.
- Compared with the standard Kratzer-style semantics, I enriched the model with a causal network, which is again necessary in capturing EXPLANATION.
- Compare with the causal premise semantics in Kaufmann (2013), I added normalcy ordering, which is the crucial ingredient that accounts for the novel observations.

Connecting NCs with EXPLANATION predicts that NCs are incompatible with semifactuals, repeated below

(27) ***Assassin's poison***:
#If not for Assassin having poisoned the coffee, Victim would sitll have died. (=8)

It also predicts that a sequence of NCs, unlike standard counterfactuals, can only occur in a certain direction.

(28) ***The shooting squad***: *There is a court, a rifleman and a prisoner. If the court orders the execution, the rifleman will shoot and the prisoner will die. The court ordered, the rifleman shot, and the prisoner died.* (Modified from Schulz 2011's example)
Q: Why did the prisoner die?

a. A: Well, if the rifleman hadn't shot, the prisoner wouldn't have died. But if the court hadn't ordered the execution, the rifleman wouldn't have shot in the first place.

b. A: Well, if not for the rifleman shooting, the prisoner wouldn't have died. But if not for the court ordering the execution, the rifleman wouldn't have shot in the first place.

(29) Q: What would happen if the court hadn't ordered the execution?

a. A: Well, if the court hadn't ordered the execution, the rifleman wouldn't have shot. And if the rifleman hadn't shot, the prisoner wouldn't have died.

b. A: #Well, if not for the court having ordered the execution, the rifleman wouldn't have shot. And if not for the rifleman having shot, the prisoner wouldn't have died.

Some more puzzles, mostly compositionality-wise:

- (30) #Victim would not have died only if not for there having been poison in the coffee.
- Morphosyntactically, what is inside *if not for*?
It might be useful to look at languages that lack an independent form for NCs. E.g. In Japanese, an NC-like interpretation seems to appear if the sententially negated antecedents contain PPIs (similar to Romero 2014's high/low negation in English counterfactuals).
- If *not* adjoins the complementizer, how to derive the effect of negation in the step of belief revision compositionally (i.e. updating the speaker's knowledge with $\neg p$ instead of $p$ in '*if not p, would q*')?

# References

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states.* The MIT press.

Halpern, J. Y. (2016). *Actual causality*. MIT Press.

Henderson, R. (2010). “if not for” counterfactuals: negating causality in natural language negating causality in natural language. In *Proceedings of the 28th West Coast Conference on Formal Linguistics*, Somerville, MA,. Cascadilla Proceedings Project.

Henderson, R. (2011). Non-defeasible counterfactuality blocks epistemic inference: Evidence from “if not for ” counterfactuals. Manuscript, UCSC.

Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *The Philosophical Review*, 116(4):495–532.

Ippolito, M. and Su, J. (2014). Counterfactuals, negation and polarity. In Crnic, L. and Sauerland, U., editors, *The Art and Craft of Semantics: A Festschrift for Irene Heim*, volume 1, MITWPL 70, pages 225–243.

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

Kratzer, A. (1981). The notional category of modality. In *Words, worlds, and contexts: New approaches in word semantics*. Walter de Gruyter.

Nevins, A. I. (2002). Counterfactuality without past tense. In Hirotani, M., editor, *Proceedings of North East Linguistic Society 32*, pages 441–450.

Pearl, J. (2000). *Causality*. Cambridge university press.

Romero, M. (2014). High negation in subjunctive conditionals and polar questions. In *Proceedings of Sinn und Bedeutung 19*, pages 499–516.

Schulz, K. (2007). *Minimal models in semantics and pragmatics: Free choice, exhaustivity, and conditionals*. PhD thesis, ILLC Dissertation Series.

Schulz, K. (2011). “if you’d wiggled a, then b would’ve changed”: Causality and counterfactual conditionals. *Synthese*, 179(2):239–251.

Sintonen, M. (1984). *The Pragmatics of Scientific Explanation*. North-Holland, Amsterdam.

Veltman, F. (1996). Defaults in update semantics. *Journal of Philosophical Logic*, 25(3):221–261.

Yang, M. (ta). Causal networks in discourse: A case of mandarin negative conditionals. In *Proceedings of North East Linguistic Society 49 (forthcoming).*