Conditionals: between language and reasoning

Class 6 - What is "similarity"?

May 28, 2019

- In minimal change semantics, the crucial semantic parameter for the evaluation of counterfactuals is a system of spheres, arising from a relative similarity ordering.
- But what is "relative similarity"?
- When does a world count as "more similar" to the actual world than another?
- If our aim is just to characterize the logic, this question can be set aside, since logical entailment is ultimately obtained by quantifying over all models.
- But the question becomes crucial if our goal is, more broadly, to understand the semantics of counterfactuals, and answer questions such as:
  - what is it about the world that makes a counterfactual true or false?
  - what processes underlie our interpretation of counterfactuals?
  - what information is conveyed by an utterance of a counterfactual?

Lewis 1973 suggests that we use our common-sensical notion of similarity.

"[Vagueness concerning the similarity of worlds] is the same sort of vagueness that arises if I say that Seattle resembles San Francisco more closely than it resembles Los Angeles. Does it? That depends on whether we attach more importance to the surrounding landscape, the architecture, the dominant industries, the political temper, [etc.]"

"Somehow, we do have a familiar notion of comparative overall similarity, even of comparative similarity of big, complicated, variegated things like whole people, whole cities, or even—I think—whole possible worlds. However mysterious that notion may be, if we analyze counterfactuals by means of it we will be left with one mystery in place of two."

Lewis 1973 suggests that we use our common-sensical notion of similarity.

"[Vagueness concerning the similarity of worlds] is the same sort of vagueness that arises if I say that Seattle resembles San Francisco more closely than it resembles Los Angeles. Does it? That depends on whether we attach more importance to the surrounding landscape, the architecture, the dominant industries, the political temper, [etc.]"

"Somehow, we do have a familiar notion of comparative overall similarity, even of comparative similarity of big, complicated, variegated things like whole people, whole cities, or even—I think—whole possible worlds. However mysterious that notion may be, if we analyze counterfactuals by means of it we will be left with one mystery in place of two."

It was soon realized, however, that assuming that the notion of similarity relevant for counterfactuals is the ordinary notion of similarity leads to wrong empirical predictions.

Fine (1975)

Context: supposed that during the Cold War, the US president had a button which would automatically unleash a large-scale nuclear attack to the Soviet Union.

(1) If Nixon had pressed the button, there would have been a nuclear holocaust.

- (1) is true.
- But now consider a world w_* which is like the actual world, except that:
  - Nixon presses the button, but the wire is cut (whereas in the actual world it isn't), so nothing happens when Nixon presses the button;
  - after pressing, Nixon comes to regret his decision and is relieved that the button failed;
  - so there is no nuclear holocaust, and history goes on roughly as we know it.
- w_* is intuitively much more similar to the actual world than any world where a nuclear holocaust takes place.
- Therefore, (1) is predicted to be false, and (2) true:

(2) If Nixon had pressed the button, the wire might have been cut.

Tichý (1976)

Jones is possessed of the following dispositions as regards wearing his hat:

- bad weather invariably induces him to wear his hat;
- fine weather affects him neither way: he wears his hat or not, at random.

Today the weather is bad, so Jones is wearing his hat.

Tichý (1976)

Jones is possessed of the following dispositions as regards wearing his hat:

- bad weather invariably induces him to wear his hat;
- fine weather affects him neither way: he wears his hat or not, at random.

Today the weather is bad, so Jones is wearing his hat.

- What if the weather were fine? Would Jones still be wearing his hat?

Tichý (1976)

Jones is possessed of the following dispositions as regards wearing his hat:

- bad weather invariably induces him to wear his hat;
- fine weather affects him neither way: he wears his hat or not, at random.

Today the weather is bad, so Jones is wearing his hat.

- What if the weather were fine? Would Jones still be wearing his hat?
- There's no telling: he might, or he might not.

Tichý (1976)

Jones is possessed of the following dispositions as regards wearing his hat:

- bad weather invariably induces him to wear his hat;
- fine weather affects him neither way: he wears his hat or not, at random.

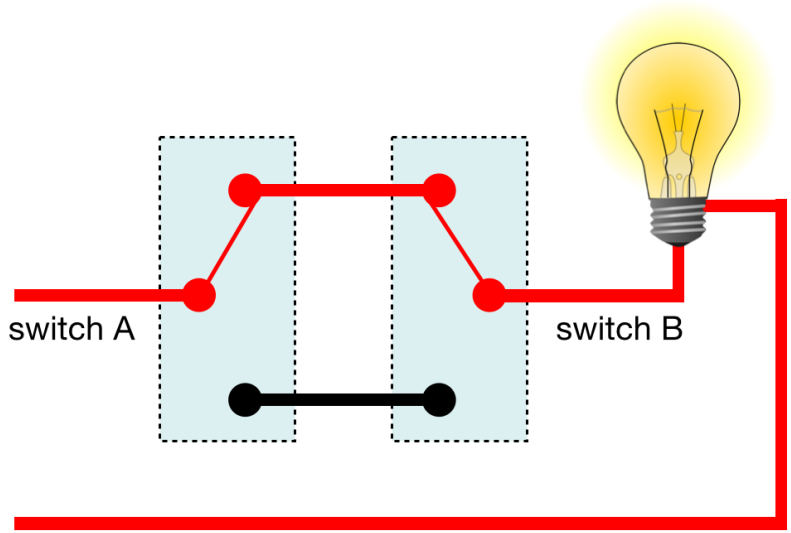
Today the weather is bad, so Jones is wearing his hat.

- What if the weather were fine? Would Jones still be wearing his hat?
- There's no telling: he might, or he might not.
- Still, worlds with good weather and hat are, intuitively speaking, more similar to actuality than worlds with good weather and no hat.
- Therefore, (3) is predicted true:

(3) If the weather were fine, Jones would still be wearing his hat.

Switches

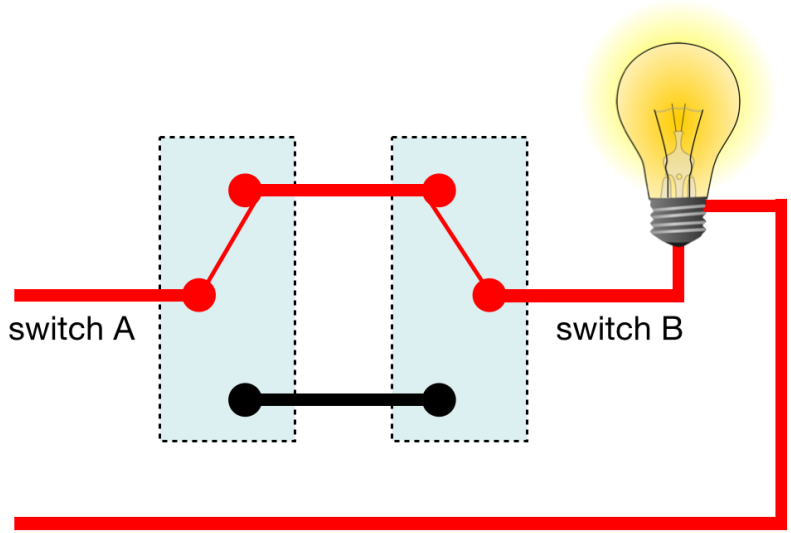
The light is on when the switches are in the same position; otherwise, the light is off. Right now, switch A and switch B are up, and the light is on.



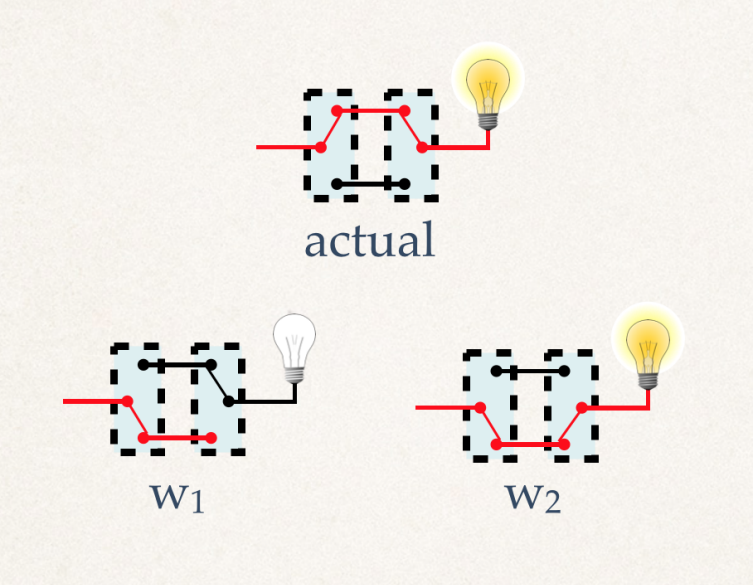
(4) If switch A was down, the light would be off.

Switches

The light is on when the switches are in the same position; otherwise, the light is off.
Right now, switch A and switch B are up, and the light is on.



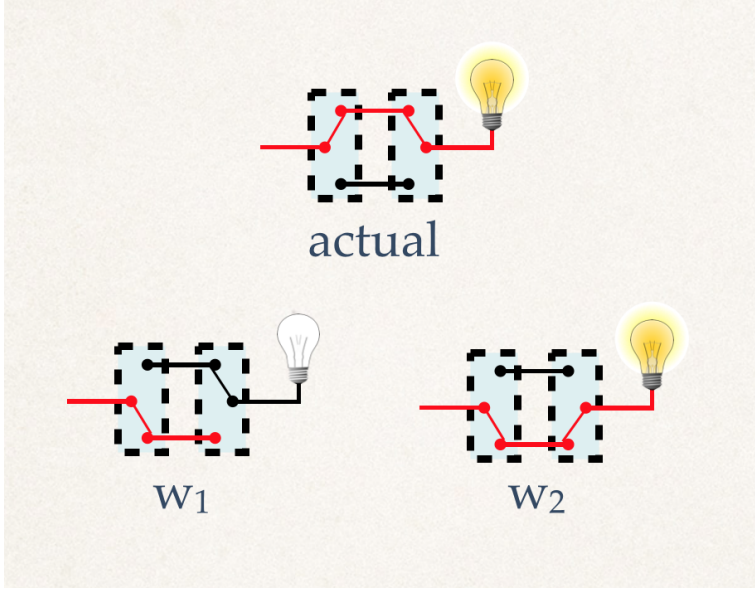
(4) If switch A was down, the light would be off. ✓



Intuitively, w_1 and w_2 seem equally similar to the actual world. In both:

- in both the laws of the circuit are obeyed;
- two facts have changed, and one is the same.

But then the counterfactual should not be true.



Intuitively, w_1 and w_2 seem equally similar to the actual world. In both:

- in both the laws of the circuit are obeyed;
- two facts have changed, and one is the same.

But then the counterfactual should not be true.

To get the right predictions, w_1 should count as more similar: changing the light should count as a smaller change than changing the position of switch B. But why?

Lewis (1979) acknowledges this

- "We will be prepared to distinguish between the similarity relations that guide our offhand explicit judgments and those that govern our counterfactuals"

- Minimal change semantics "is only a skeleton. It must be fleshed out with an account of the appropriate similarity relation, and this will differ from context to context. Our present task is to see what sort of similarity relation can be combined with [it] to yield [. . .] the standard resolution of vagueness."

- By a "standard resolution of vagueness" Lewis means the one which underlies our ordinary judgments of counterfactuals, corresponding to causal relationships—according to which (5) is true and (6) false.

  (5) If N. had pressed the button, there would have been a nuclear holocaust.

  (6) If N. had pressed the button, the wire would have been cut.

- Lewis uses the Nixon example to come up with criteria to determine the relevant similarity relation. Let's follow his reasoning.

- Let w_0 be the relevant world and t be the relevant time.

- Suppose w_0 obeys deterministic laws and "fits our worst fantasies about the button: there is such a button, it is connected to a fully automatic command and control stystem, the wired-in war plan consists of one big salvo, everything is in faultless working order, there is no way for anyone to stop the attack, and so on."

- Then Fine's counterfactual is true at w_0:

  (7) If Nixon had pressed the button, there would have been a nuclear holocaust.

"There are all sorts of worlds where Nixon presses the button at t. We must consider which of these differ least, under the appropriate similarity relation, from w_0. [...] The more serious candidates fall into several classes."

Type 1

- Everything is exactly the same as in w_0 until shortly before t.
- Around t, a tiny miracle occurs which leads to Nixon pressing the button.
- By "miracle" Lewis means an exception to the deterministic laws of w_0 (not of w_1; Lewis assumes that laws are at least true regularities).
- After that, things proceed in accordance with the laws of w_0.
- As a result a nuclear holocaust takes place.
- These are the worlds that should turn out to be the closest to w_0.

Type 2

- In these worlds, we have no miracles: the laws of w_0 are obeyed exactly.
- Since laws are deterministic, such worlds must differ from w_0 at all times.
- If these worlds were the closest, many backtracking counterfactuals would be true:

(8) If Nixon had pressed the button, Napoleon might never have existed.

- Under the "standard resolution", such counterfactuals are not true.
- So, worlds of type 2 should be further away than those of type 1.
- Exact match of facts for more time counts more than avoidance of small miracles.

Type 3

- Everything is exactly the same as in w_0 until shortly before t.
- Around t, a small miracle occurs, leading to Nixon pressing the button.
- After t, a small miracle occurs, preventing the holocaust: say, the wire breaks.
- Things proceed in accordance with the laws (of w_0).
- No nuclear holocaust occurs; things are approximately similar to w_0.
- But not exactly the same; many small differences remain: Nixon remembers what happened, the wire is broken, the click of the button is recorded on a tape...
- We gained approximate match of facts at the price of a small extra miracle.
- To get the predictions we want, these worlds should be further away than type 1.
- So, avoiding small miracles is more important than having approximate match of facts for a longer time.

Type 4

- Everything is exactly the same as in w_0 until shortly before t.
- Around t, a small miracle occurs, leading to Nixon pressing the button.
- After t, a large and widespread miracle occurs: the wire breaks and fixes itself immediately; Nixon's memories are falsified; ...
- The following course of history matches perfectly the one of w_0.
- Now we gained exact match of facts at the price of a large miracle.
- Again, such worlds should be further away than those of type 1.
- So avoiding large miracles is more important than having perfect match of facts for a longer time.

Lewis concludes that, for the standard reading of counterfactuals, relative similarity should be governed by the following system of priorities.

1. It is of the first importance to avoid big, widespread violations of law.

2. It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.

3. It is of the third importance to avoid even small, localized, simple violations of law.

4. It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

Conceptual concerns

- These priorities look badly *ad hoc*: even if they were empirically adequate, one would like some explanation for why counterfactuals are interpreted this way.
- Why would our hypothetical reasoning be driven by such a strange system of priorities?
- E.g., why would exact match of facts matter more than small violations of laws but less than large ones?

Empirical problem I

"It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails."

Coat thieves (Nute 1980)

- I forgot my coat in the bar last night.
- In the course of the night two potential coat thieves passed by the coat, one at 10, the other at midnight.
- Each time, there was a non-zero chance that the coat would be stolen.
- The next morning I find to my relief that the coat is still where I left it.

Empirical problem I

"It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails."

Coat thieves (Nute 1980)

- I forgot my coat in the bar last night.
- In the course of the night two potential coat thieves passed by the coat, one at 10, the other at midnight.
- Each time, there was a non-zero chance that the coat would be stolen.
- The next morning I find to my relief that the coat is still where I left it.

(9) If the coat had been stolen, it would have been stolen at midnight.

Empirical problem I

"It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails."

Coat thieves (Nute 1980)

- I forgot my coat in the bar last night.
- In the course of the night two potential coat thieves passed by the coat, one at 10, the other at midnight.
- Each time, there was a non-zero chance that the coat would be stolen.
- The next morning I find to my relief that the coat is still where I left it.

(9) If the coat had been stolen, it would have been stolen at midnight.

- Not necessarily — it might just as well have been stolen at 10.

- Yet consider Lewis's system of priorities.
- Consider any two worlds w_1 and w_2 which are like the actual world except that a small miracle makes, respectively, the first and the second thief steal the coat.
- Priority 1 is satisfied by both: in both cases no large miracle is required.
- Priority 2 asks to maximize spatio-temporal region where we have exact match: w_2 wins here, since the departure from the actual world happens later.
- So, the closest antecedent worlds are those where the theft happens at midnight.
- So (10) is predicted to be true:

(10) If the coat had been stolen, it would have been stolen at midnight.

Empirical problem II

"It is of little or no importance to secure approximate similarity of particular fact."

- Little importance, or no importance?
- Either way, it seems that wrong predictions are made.

Jones, version 1

Every morning Jones looks out of the window. If the weather is bad, he wears his hat. If it is good, he flips a coin to decide whether to wear his hat (heads$\rightsquigarrow$yes; tails$\rightsquigarrow$no).

Today, the weather is bad, so Jones is wearing his hat.

(11) If the weather were fine, Jones would still be wearing his hat.

Jones, version 1

Every morning Jones looks out of the window. If the weather is bad, he wears his hat. If it is good, he flips a coin to decide whether to wear his hat (heads$\rightsquigarrow$yes; tails$\rightsquigarrow$no).

Today, the weather is bad, so Jones is wearing his hat.

(11) If the weather were fine, Jones would still be wearing his hat. ??

Jones, version 1

Every morning Jones looks out of the window. If the weather is bad, he wears his hat. If it is good, he flips a coin to decide whether to wear his hat (heads$\rightsquigarrow$yes; tails$\rightsquigarrow$no).

Today, the weather is bad, so Jones is wearing his hat.

(11) If the weather were fine, Jones would still be wearing his hat. ??

Jones, version 2 (Veltman 2005)

Every morning Jones flips a coin, then looks out the window. If the weather is bad, he wears his hat. If it's good, he decides based on the coin (heads$\rightsquigarrow$yes; tails$\rightsquigarrow$no).

Today, the coin came down heads, and the weather is bad, so Jones is wearing his hat.

(12) If the weather were fine, Jones would still be wearing his hat.

Jones, version 1

Every morning Jones looks out of the window. If the weather is bad, he wears his hat. If it is good, he flips a coin to decide whether to wear his hat (heads$\rightsquigarrow$yes; tails$\rightsquigarrow$no).

Today, the weather is bad, so Jones is wearing his hat.

(11) If the weather were fine, Jones would still be wearing his hat. ??

Jones, version 2 (Veltman 2005)

Every morning Jones flips a coin, then looks out the window. If the weather is bad, he wears his hat. If it's good, he decides based on the coin (heads$\rightsquigarrow$yes; tails$\rightsquigarrow$no).

Today, the coin came down heads, and the weather is bad, so Jones is wearing his hat.

(12) If the weather were fine, Jones would still be wearing his hat. ✓

Lewis's system of weights does not explain the contrast

- Let w_h and w_t be like the actual world, except that a small miracle occurs during the night which makes the weather good. Moreover:
  - In w_h the coin comes down heads and Jones wears his hat.
  - In w_t the coin comes down tails and he doesn't.

- w_h and w_t are the same with respect to criteria 1 (avoid large miracles), 2 (maximize exact match of history), and 3 (avoid small miracles).

- In both scenarios w_h is closer than w_t in terms of approximate match of facts.

- Does approximate match of facts count for something?
  - Yes: then w_h is closer than w_t; so (13) is predicted true.
  - No: then w_h and w_t are equally close; so (13) is not predicted true.

  (13) If the weather were fine, Jones would still be wearing his hat.

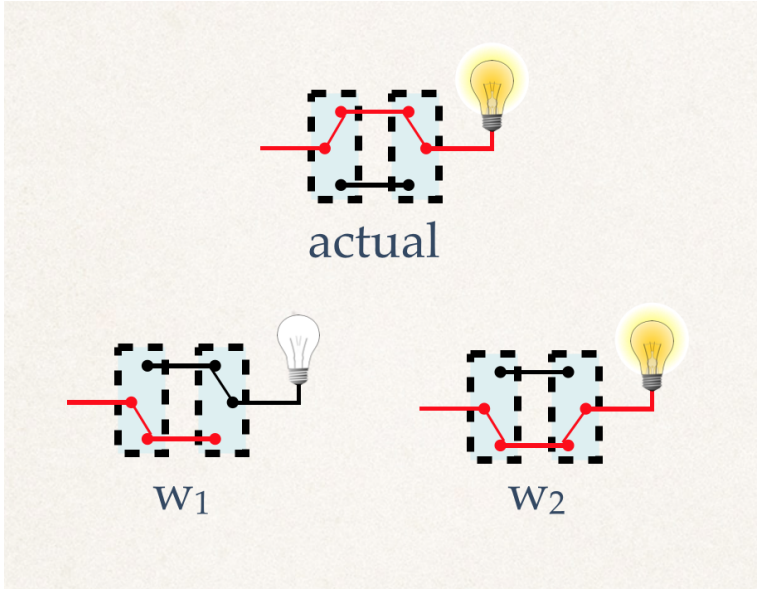
- Either way, the same predictions are made in the two cases.

- So, Lewis's recipe for determining similarity is not satisfactory.
- One may attack the problem of similarity in a different way, by giving a different theory of how similarity is determined.
- As a matter of fact, however, the most elegant answers to our questions came from theories that do not start out with the Lewisian picture at all.
- (One may still ask if these theories can be identified with minimal change semantics under a certain theory of similarity; the answer is: to some extent.)
- In order to see the main idea of these theories, let us ask: what is the key difference between the two Jones' cases?

Veltman's diagnosis of the contrast

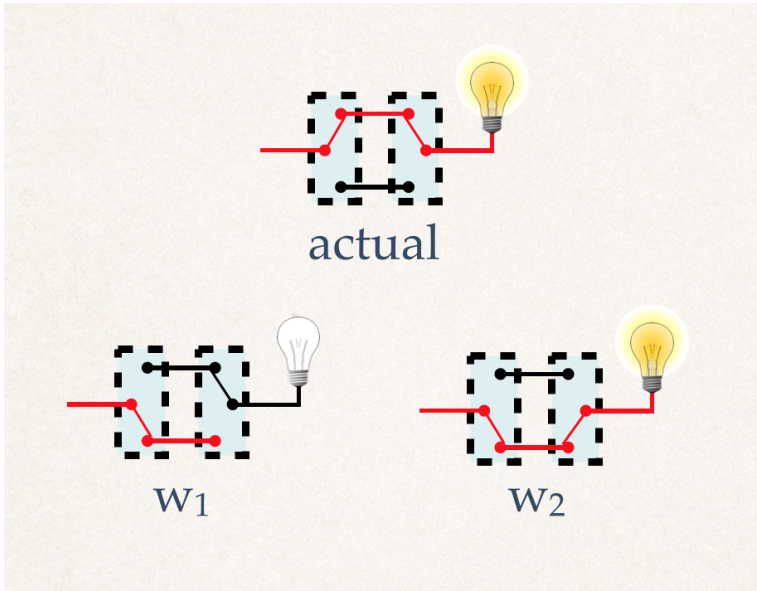
- In both cases Jones is wearing his hat because the weather is bad.
- In both cases we have to give up the proposition that the weather is bad —the very reason why Jones is wearing his hat.
- So, why should we want to keep assuming that he has his hat on?
- In the first case there is no special reason to do so; hence, we do not.
- In the second case there is a special reason. We will keep assuming that Jones is wearing his hat because we do not want to give up the independent information that the coin came down heads.
- And this, together with the counterfactual assumption that the weather is fine, brings in its train that Jones would have been wearing his hat.
- In sum: in making a counterfactual assumption, we only strive to hold fixed those facts about the world which are independent of the assumption we are making.

Switches



Making the assumption that A is down, leads us to consider the world w_1. Why?

Switches



Making the assumption that A is down, leads us to consider the world w_1. Why?

- The state of the light is dependent on the position of switch A.
- The position of switch B is independent of the position of switch A.

Crucial insight

- When we evaluate counterfactuals, we view the world not as a mere set of facts, but as a network in which facts are connected by (causal) dependency relations.
- The semantics of counterfactuals is highly sensitive to this network structure.
- This important idea is at the core of causal accounts of counterfactuals.
- These approaches give us a much more concrete, operational grip on the process of making counterfactual assumptions and computing their consequences.