# Conditional Logic and Belief Revision

Ginger Schultheis (vks@mit.edu) and David Boylan (dboylan@mit.edu)

January 2017

## *What are conditionals?*

▷ Take sentences like:

(1) If I hadn't slept in, I would have made it in time.

(2) If the coin had come up heads, I would have won twenty dollars.

These are *conditionals*. Today we'll be discussing theories of what it takes to make a conditional true and some issues such theories raise.

▷ Quickly let's note the indicative/subjunctive distinction:

(3) If Oswald didn't shoot Kennedy, somebody else did. (indicative)

(4) If Oswald hadn't shot Kennedy, somebody else would have. (subjunctive)

▷ Roughly speaking, subjunctives (also called *counterfactuals*) ask what would have happened if something had gone differently.

Indicatives ask what will follow if something (we don't know yet know) turns out to be true.

This should help you get a feel for the difference but it isn't exactly right. Can you think of counterexamples to each of these claims?

▷ These examples show there must be some difference in meaning between indicatives and subjunctives: after all, one is true and the other is dubious.

Today we will focus on counterfactuals.

## *What is belief revision?*

Suppose you believe the following:

▷ The bird caught in the trap is a swan. ($A$)

▷ The bird caught in the trap comes from Sweden. ($B$)

▷ Sweden is part of Europe. ($B$)

▷ All European swans are white. $D$

If you believe all of these things, it follows that you *also* believe:

▷ The bird caught in the trap is white (E).

Now suppose you're told, by a reliable source, that the bird caught in the trap is not white. Now you want to add $\neg E$ to your beliefs. But you can't *just* do that—that would make your belief set inconsistent. So you need to *subtract* some other beliefs. **Belief revision is the process of adding a belief and subtracting others to ensure consistency**.

It's implicit that the only reason to subtract a belief from belief set $B$ is that you've learned something *inconsistent* with $B$. There's an axiom that corresponds to this implicit assumption, called **Preservation**. We'll return to this later.

▷ There are two parts to the study of belief revision: the first is stating general principles that govern belief change; the second is constructing a belief revision operation that satisfies those postulates.

- Some questions to think about when studying belief revision: How are the beliefs in a dataset represented? What is the relationship between the elements explictly represented in the database and those that are derived from the ones explicitly represented? How are choices concerning what to retract made?

Our answers to these questions may well depend on the application area.

## *Why study these things together?*

- Ramsey's test

> According to the suggestion, your deliberation .... should consist of a simple thought experiment: add the antecedent (hypothetically) to your stock of knowledge (or beliefs), and then consider whether or not the consequent is true. Your belief about the conditional should be the same as your hypothetical belief, under this condition, about the consequent (Stalnaker, 1968).

To believe a conditional just is to form a (hypothetical) policy of belief revision.

- Formal machinery
  - Models developed for studying conditionals can also be used to study belief revision.
- Strengthening principles and defeasible reasoning
  - The conditional $p > q$ says that, in *some* sense, $q$ follows from $p$; from $p$ one can infer $q$. This inference is not *deductively valid*, since it seems to be *defeasible*:

(1) If I had gone to the party, I would have had fun.

(2) If I had gone to the party and spotted my ex with someone new, I would not have had fun.

In general, one cannot add 'premises' to the antecedent and expect that the consequent will still follow—*stregthening the antecedent* is not valid for conditionals.

- (1) and (2) may both be true.
- But perhaps some restricted strengthening principles are valid. Suppose that both (3) and (4) are true:

(3) If you had gone to the party, you would've had fun.

(4) If you had gone to the party, you would have had a few beers.

It seems to follow that,

(5) If you had gone to the party and had a few beers, I would have fun.

We can add a 'premise' to the antecedent of a conditional when that premise *itself* counterfactually follows from the antecedent

- Beliefs are also defeasible. If you tell me you went to a party last night, I'll assume you had fun. If you later tell me you spotted your ex with someone new, I'll no longer believe you had fun.
  - In the lingo: That you $q$ conditional on $p$ does not entail that you believe $q$ conditional on ($p$ and $r$).

That you spotted your ex *defeats* my belief that you had fun at the party.

- But we can still ask, as we did for conditionals, whether certain restricted 'strengthening principles' are valid for conditional belief.

## *What's our question about counterfactuals?*

- Take a counterfactual like (4). Our question is, what generally does it take for counterfactuals to be true?
- To get a grip on this question, compare it to a similar but simpler one: what does it take for

This question is important as giving a theory of the meanings of expressions must yield (at least) their truth-conditions.

(5) It's not raining.

to be true? We'll it just has to be false that it's raining. More generally, what it takes for $\neg A$ to be true is for $A$ to be false.

▷ This is an important point: we can analyse the sentence as having two main components: 'It's raining' and 'not' and we see that the truth-value of the resulting sentence is completely determined by the truth-value of the embedded sentence.

▷ Counterfactuals are hard because they don't just depend on the truth values of the antecedent and consequent.

Exercise: imagine two different worlds in both of which Oswald actually shoots Kennedy (and no body else does), but where only one makes the counterfactual true.

▷ What seems relevant is not just what our world is like, but what other similar worlds are be like. In particular, we want to ask what would happen in a world where things are almost the same, except for the fact that the antecedent is true there.

The theory about counterfactuals we'll talk about tries to make this intuition precise.

▷ We'll do this by using methods in formal logic and semantics:

- We'll introduce a formal language with a connective $>$ that will capture the conditional.
- We'll give a possible worlds model for the language which will tells us in which worlds the conditional is true.

## *Basic Similarity Semantics for Conditionals*

Here is the guiding idea for the semantics:

> Consider a possibile world in which $A$ is true, but otherwise differs minimally from the actual world. The conditional *If A then B* is true (false) just in case $B$ is true (false) in that possible world (Lewis, 1973).

Call this the *variably strict* analysis of conditionals.

Basic Semantics

▷ Language: we have a formal language with atomic sentences $p$, $q$..., $\wedge$, $\neg$, $\supset$ and $>$.

Set of sentences is defined recursively in the usual way.

▷ Models are triples $\langle W, \leq, V \rangle$

- $W$ is a set of points; think of these as *possible worlds*, various different ways the world could be.
- $\leq$ takes a world $w$ and returns an ordering on worlds $\leq_w$; understand $w_1 \leq_w w_2$ as saying that $w_1$ is at least as similar to $w$ as $w_2$ is.
- $V$ is a function which takes atomic formula and worlds and returns a truth-value.

For now, don't assume any of the intuitive properties of similarity; we will be adding those features to the model one by one to observe the validities they give rise to.

▷ We need now to say when sentences are true at worlds in in our model. We do this recursively:

- $\mathcal{M}, w \vDash p$ just in case $V(p, w) = 1$
- $\mathcal{M}, w \vDash \neg\phi$ iff not $\mathcal{M}, w \vDash \phi$

Also $\mathcal{M}, w \vDash \phi \wedge \psi$ iff $\mathcal{M}, w \vDash \phi$ and $\mathcal{M}, w \vDash \psi$ and $\mathcal{M}, w \vDash \phi \supset \psi$ iff not $\mathcal{M}, w \vDash \phi$ and $\mathcal{M}, w \vDash \neg\psi$

▷ First try at the counterfactual: Say that $f(\phi, w)$ is the set of worlds in $[\![\phi]\!]$ which are closest to $w$ (i.e. the closest $\phi$ worlds). Now say:

$$\mathcal{M}, w \vDash \phi > \psi \text{ iff } \forall w' \in f(\phi, w) : \mathcal{M}, w' \vDash \psi$$

In English: $\phi > \psi$ is true just in case $\psi$ is true in all the closest $\phi$-worlds.

▷ However, what happens if we have infinitely many worlds?
Let's try again:

$\mathcal{M}, w \vDash \phi > \psi$ iff for every $w_1 \in [\![\phi]\!]$:

- $\exists w_2$ s.t. $w_2 \leq_w w_1$ and $w_2 \in [\![\phi]\!]$;
- and $\forall w_3 \in [\![\phi]\!]$ s.t. $w_3 = w_2$ or $w_3 < w_2 : \mathcal{M}, w_3 \vDash \psi$

▷ Equivalently: $\phi > \psi$ is true just in case some $\phi$ and $\psi$-world is closer than any $\phi$ and $\neg\psi$ world.

▷ With no further constraints on the similarity relation, this gives us the following weak logic:

Taut    If $\phi$ is a classical tautology, then $\vdash \phi$.

CI    $\vdash \phi > \phi$

CC    $\vdash (\phi > \psi) \wedge (\phi > \chi) > (\phi > (\psi \wedge \chi))$

CW    $\vdash (\phi > \psi) > (\phi > (\psi \vee \chi))$

ASC    $\vdash (\phi > \psi) \wedge (\phi > \chi) > (\phi \wedge \psi > \chi)$

AD    $\vdash (\phi > \chi) \wedge (\psi > \chi) > ((\phi \vee \psi) > \chi)$

MP⊃    $\phi \supset \psi, \phi \vdash \psi$

Note that we do *not* yet have MP for >.

REA    If $\vdash \phi \equiv \psi$, then $\phi > \chi \vdash \psi > \chi$

## Adding Conditions

> is a kind of modal operator and so you can think of ≤ as yielding a kind of accessibility relation. Just as in normal modal logic, imposing certain properties on ≤ will result in validating certain principles.

▷ Start with the property *Weak Centering*:

WC    For all $w$, $\neg\exists w'$ s.t. $w' \leq_w w$ and not $w \leq_w w'$.

In English: no worlds is more similar to $w$ than $w$ is to itself. (This seems non-negotiable for similarity.)

▷ Notice that if WC holds we validate the following inference:

MP>    $\phi \wedge (\phi > \psi) \vdash \psi$

Q.1. Can you show why this is?

Q.2 Does this inference seem right to you? Can you give examples to support your answer?

▷ Here is a similar property, *Strong Centering*:

SC For all $w, w'$, if $w' \leq_w$ then $w' = w$.

In English: $w$ is closer to itself than anything else. This also seems reasonable for similarity.

▷ SC gives us the following inference:

TT $\phi, \psi \vdash \phi > \psi$

Q.1 Can you show why this is?

Q.2 Does this kind of inference seem good to you?

▷ Here's a controversial property, the Limit Assumption:

LA For every set of worlds $S$, $\exists w' \in S$ s.t. $\neg\exists w'' \in S : w'' <_w w'$.

In English: take any set of worlds and you'll find there's at least one world in that set which is as close to $w$ as any other.

Equivalently: take any set of worlds and you'll find there's at least one closest world to $w$.

▷ If we have LA, then we can state our semantics in the much simpler way

$>$ $\mathcal{M}, w \vDash \phi > \psi$ iff $\forall w' \in f(\phi, w) : \mathcal{M}, w' \vDash \psi$

This is because LA guarantees that, even in infinite models, there will always be some best $\phi$ world.

▷ Here's a principle called connectedness:

Con $\forall w_1, w_2$: $w_1 <_w w_2$, $w_2 <_w w_2$ or $w_1 = w_2$.

Equivalently: for any two distinct worlds, one is closer to $w$ than the other.

▷ Con and LA together give us the principle *Conditional Excluded Middle*:

CEM $(\phi > \psi) \vee (\phi > \neg\psi)$

Q.1 Why do these properties give us CEM?

Q.2 Is CEM plausible? Can you think of examples to support your answer?

## *Questions about Similarity*

We've said that the counterfactual *if A had been the case, B would have been the case* is true just all of the most similar $A$-worlds are also $B$-worlds. But which $A$-worlds count as the most similar $A$-worlds? Here are some things we'd like to say more about.

- We're not interested in worlds that differ from ours *only* in that the the antecedent is false. Consider a world where I go to the beach today. Is my car still parked at home? Am I still wearing my slippers? Am I still teaching this course?
- Determining how similar *overall* one world is to another requires weighing respects of similarity. But how do we do that? Here's an example that makes this particularly pressing:

  > Imagine there is a button that launches the entire US nuclear arsenal. Nixon is standing by. He decides not to press the button. It seems true that
  >
  > (6) If Nixon had pushed the button, there would have been a nuclear catastrophe.
  >
  > If (6) is true, then a world where Nixon pushes the button and there's nuclear is more similar to the actual world than one where he pushes the button and there is no nuclear catastrophe. But is it?

- Similarity is a context-sensitive affair. Consider:

  (7) If Ceasar had been in command in Korea, he would have used the A-bomb.

  (8) If Ceasar had been in command in Korea, he would have used catapults.

  (7) and (8) can both seem true depending on the context in which they're uttered.

In different contexts, we hold different facts fixed, e.g., that Ceasar does or does not have access to modern-day weapons, that he does or does not have knowledge modern-day weapons, etc.

- Is there always a *single* closest *A*-world?
  - Suppose David and I are sitting on top of each other in the same chair right this moment. Is he sitting on top of me or am I on top of him? Perhaps these worlds are *equally close* to the actual world.

The assumption that there is always a *single* closest antecedent-world is the **Uniqueness Assumption.**

- Is there always *some* closest A-world?
  - Lewis says no:

    > Just as there is no shortest possible legnth above one, so there is no closest world to ours among the worlds with lines more than an inch long (Lewis, 1973).

The assumption that there is always *some* closest A-world (or set of closest A-worlds) is the **Limit Assumption**.

- Clearly, the actual world is *among* the closest worlds to actuality. But could there be other worlds that are *just as* close to the actual world as the actual world is to itself?

The assumption that the actual world is among the closest worlds to actuality is **Weak Centering**. The assumption that the *single closest* world to actuality is **Strong Centering**.

## *Strengthening Principles*

- What is a strengthening principle? We'll take a strengthening principle to be something of the form

  (6) $\phi > \chi \wedge \alpha_1 \wedge ... \wedge \alpha_n \supset \phi \wedge \psi > \chi$

  That is, it's a principle which allows us to go from the premise that if $\phi$ had happened $\chi$ would have happened, together with certain assumptions, to the conclusion that if $\phi$ and $\psi$ had happened, $\chi$ would still have happened.

As we said in the beginning, these principles are interesting because counterfactual reasoning in *defeasible* in a certain kind of way. Different strengthening principles will tell us under which kinds of assumptions make these inferences safe ones to make.

▷ Note that we can order these principles according to which ones entail others. We say that one principle is stronger than another if it entails it (but not vice versa).

We're going to talk about three today (in increasing order of strength): ASC, RM and AS.

It's worth asking yourself whether you can think of other interesting strengthening principles.

▷ **Antecedent Strengthening with the Consequent**

- Recall

ASC $\vdash (\phi > \psi) \wedge (\phi > \chi) \supset (\phi \wedge \psi > \chi)$

- ASC looks hard to dispute: if both

(7) If I had woken up late, I would have missed the bus.

(8) If I had woken up late, I would have forgotten my backpack.

then it seems to follow that

(9) If I had woken up late and missed the bus, I would have forgotten my backpack.

- ASC is somewhat interesting because it might be the weakest strengthening principle which is definitely true.
- Open question: can you think of a stronger principle (i.e. one that entails ASC) but which is also clearly true?

▷ **Antecedent Strengthening**

- Antecedent strengthening is this principle:

AS $(\phi > \chi) \supset ((\phi \wedge \psi) > \chi)$

- This principle has an important place in the debate on conditionals. Invalidating it was a major motivation for developing the variably strict analysis and to move away from a *strict* semantics for the counterfactual.
- The strict semantics is simple: take the material conditional and put a necessity over it:
  - ▷ $B(w) = \{w' : \neg\exists w'' : w'' <_w w'\}$
  - ▷ $\mathcal{M}, w \vDash \Box\phi$ iff $\forall w' \in B(w) : \mathcal{M}, w' \vDash \phi$.
  - ▷ $\mathcal{M}, w \vDash \Box(\phi \supset \psi)$ iff $\forall w' \in B(w) : \mathcal{M}, w' \vDash \phi \supset \psi$.
- AS is valid on a strict semantics, but not on a variably strict semantics.

This initially looks good for variably strict analysis. Consider the following sequence:

(10) a. If I had struck the match, it would have lit.

b. But if I had struck the match and it had been soaked in water overnight, it would not have lit.

This looks like a counterexample to AS: if the first sentence is true, then the second should be false. (Why is this?)

- Lately there has been some push back to this argument. Consider this:

(11) a. If I had struck the match and it had been soaked in water overnight, it would not have lit.
b. But if I had struck the match, it would have lit.

The sequence no longer sounds so good when it's reversed. This is puzzling for the variably strict account.

▷ **Rational Monotonicity/Strengthening with a Possibility**

- RM $(\phi > \chi) \wedge (\neg\phi > \neg\chi) \supset (\phi \wedge \psi > \chi)$
  This principle is more famous in its belief-revision guise (more on that tomorrow). But it raises interesting questions for counterfactuals too.
- The principle which gives us RM is called *almost-connectedness*:

AC $\forall w_1, w_2, w_3$: if $w_1 < w_2$, then $w_1 < w_3$ or $w_2 < w_3$

Q. Why does AC give us RM?

- (Potential) counterexample: Ginsberg claims this inference is invalid.

If Verdi and Satie had been compatriots, Bizet would have been French.
It's not the case that if Verdi and Satie had been compatriots, Satie and Bizet would not have been compatriots.
Therefore if Verdi and Satie and Bizet had been compatriots, Bizet would have been compatriots.

What do you think about this inference?

- Final note: RM is entailed by AS and so has potential upshots for the debate about AS. If it turns out the techniques to explain counterexamples to AS do not apply to counterexamples to RM, this is a problem for the strict conditional view.