Counterfactuals and Belief Revision

Marcello Di Bello
University of Amsterdam

August 22, 2007

Today's Plan

1. Counterfactuals
   D. Lewis (1973), *Counterfactuals*, Oxford UP.

2. Belief revision
   K. Segerberg (1998), 'Irrevocable Belief Revision in Dynamic Doxastic Logic', NDJFL.

Plan

Counterfactuals

Counterfactual Sentences

- If Uribe were a honest man, he would say that he is a murder.
- If Al Gore had won the US election in 2000, the US would not have occupied Iraq in March 2003.
- If philosophy were not taught at the university, some money would be used in a different way (e.g., given to poor people).

If it **were** the case that φ, then it **would be** the case that ψ.

$\varphi \rightsquigarrow \psi$.

On the Methodology in Formal Semantics

Fact 1 Speakers of a language can tell whether a sentence is true or not; or they can tell in which circumstances a sentence is true or false.

Fact 2 We seek a formal theory about counterfactual sentences.

Goal Thus, we seek a formal theory that agrees with the speakers' intuitions about the truth-value of counterfactuals sentences.

First Attempt: Material Implication

Proposal $\varphi \rightsquigarrow \psi$ iff $\varphi \rightarrow \psi$.

Problem Antecedents of counterfactual sentences are typically false. Thus, every counterfactual sentence would be vacuously true. This is counterintuitive.

Second Attempt: Strict Implication

Proposal $w \Vdash \varphi \rightsquigarrow \psi$ iff $w \Vdash \Box(\varphi \rightarrow \psi)$.

Question How do we define the accessibility relation R_w?
$R_w = \{(w, v) : v \text{ is similar to } w\}$.

Remark Intuitively, the R_w-accessible worlds should be those in which φ is false, but everything else is the same as in w. But this is not possible (why?). Thus, we need a notion of similarity.

Problem In counterfactual reasoning monotonicity does not hold.

Aside: Two Equivalent Notations

N1 Given a world w, the R_w-**accessible worlds** are those that are similar (according to a certain degree of similarity) to w.
$R_w = \{(w, v), (w, u), \dots\}$

N2 Given a world w, we can define a **sphere of worlds around** w. This is the sphere of worlds that are similar to w (according to a certain degree of similarity).

Failure of Monotonicity and Counterfactuality

(1) If I were a workaholic, my wife would complain that I never pay attention to her.

(2) If I were a workaholic *and my wife were dead*, my wife would complain that I never pay attention to her.

Fact Sentence (1) is true (or assume it is true). Sentence (2) is clearly false.

Problem Under the strict implication solution, if sentence (1) is true, then sentence (2) is true as well. $\Box(\varphi \rightarrow \psi) \rightarrow \Box(\varphi \wedge \chi \rightarrow \psi)$ is a valid formula.

Upshot It seems we need to change the R_w-accesible worlds, depending on the antecedent of the counterfactual sentence. The degree of similarity to be considered varies. Thus, Lewis proposes to see counterfactuals as **varying strict conditionals**.

Solution: Systems of Spheres $\mathcal{S}_i$

Let $\mathcal{W}$ be the logical space. Let $\wp(\mathcal{W})$ be the power set of $\mathcal{W}$. Let w be the actual world (or the world we want to evaluate the counterfactual from).

A system of spheres $\mathcal{S}_w \subseteq \wp(\mathcal{W})$ satisfies:

- CENT: $\{w\} \in \mathcal{S}_w$;
- NEST: for any $X, Y \in \mathcal{S}_w$, we have $X \subseteq Y$ or $Y \subseteq X$;
- C-UN: $\bigcup C \in \mathcal{S}_w$, for any $C \subseteq \mathcal{S}_w$;
- C-IN: $\bigcap C \in \mathcal{S}_w$, for any $C \subseteq \mathcal{S}_w$.

Notice that if $\mathcal{S}_w$ is finite, then C-UN and C-IN follow from NEST.

Example: Systems of Spheres

$\mathcal{S}_{w_1} = \{\{w_1, u\}, \{w_1, u, v\}\}$ **no!**
$\mathcal{S}_{w_2} = \{\{w_2\}\{w_2, u\}, \{u, v\}\}$ **no!**
$\mathcal{S}_{w_3} = \{\{w_3\}, \{w_3, u\}, \{w_3, u, v\}\}$ **yes!**

- w is more similar to w than u.
- u is more similar to w than v.
- Thus, $R_w = \{(w, w), (w, u), (u, v), (w, v)\}$.

Truth-conditions Based on Systems of Spheres $\mathcal{S}_i$ (First Formulation)

Solution $w \Vdash \varphi \rightsquigarrow \psi$ is true iff

(i) no φ-world belongs to any sphere in $\mathcal{S}_w$.

(ii) there is a sphere in $\mathcal{S}_w$ such that:
it contains some φ-worlds; and
$\varphi \rightarrow \psi$ is true in any world in that sphere.

Claim: The system of spheres solution solves the problem with monotonicity.

Second Formulation: The Limit Condition

Sometimes it is customary to spell out the truth-conditions for counterfactuals this way:

Truth-conditions $w \Vdash \varphi \rightsquigarrow \psi$ is true iff

(i) no φ-world belongs to any sphere in $\mathcal{S}_w$.

(ii) in the smallest φ-sphere, the formula $\varphi \rightarrow \psi$ is true.

L-Ass: Given an subset $X \subseteq \mathcal{S}_w$, there is a minimal element $M \in X$, i.e., if $M' \in X$, then $M' \subseteq M$.

L-Ass: Given an subset $X \subseteq \mathcal{S}_w$, we have $\bigcap X \in X$.

The Limit Condition is Problematic

Consider

(1) If Edgar were shorter than he is, he would not ....

Suppose Edgar is n meters tall. Thus, there are worlds in which Edgar is $n-1$ meters tall, worlds in which he is $n-\frac{1}{2}$ meters tall, worlds in which he is $n-\frac{1}{3}$ meters tall, and so on. This would give rise to an *infinite descending chain* of spheres of worlds, thus invalidating the limit condition.

Remark: Everything depends on how we set up the similarity relation between worlds. For instance, we could say that worlds in which Edgar is $n-\frac{1}{2}$ and worlds in which he is $n-\frac{1}{3}$ have the same degree of similarity. But, then, how do we decide about degrees of similarity?

A Problematic Counterfactual Sentence

Scenario: Ben bets tails. Alice flips the coin, and it lands heads. The coin toss is fair and indeterministic.

Now the following counterfactual should be true:

(1) If Ben had bet heads, he would have won.

Consider the similar (counterfactual) worlds, i.e., the ones in which Ben bets tail. On which ground should we suppose that in **all** these worlds the coin lands heads? In some of the counterfactual worlds, the coin will land tails, and in some others it will land heads. Thus, (1) should be false.

Counterfactual and Causality

- The problem with sentence (1) have induced some to abandon Lewis' analysis and suggest that counterfactuals should be analysed in terms of **casual models**.
- Conversely, Lewis intends to use his analysis of counterfactuals to give a reductive account of causality.
- Which notion comes first: causality or counterfactuality?

Plan

Belief revision

Belief Change

- Artificial or natural agents are endowed with a set of beliefs (opinions about how the world is like).
- Agents develop and modify their belief sets, depending on the new information they are confronted with.
- **If** an agent accepts the new information she has come across, **then** she can
  1. **extend** her belief set by adding the new information (the new information is *consistent* with the old belief set).
  2. **revise** her belief set (the new information is *inconsistent* with the old belief set). (revision can be though if as the double operation of **contraction** plus **extension**.)

Belief Revision: Example 1

Suppose this is (part of) what an agent believes:

- B1 All landlords are good people.
- B2 The man I saw in Plaza de Bolivar is a landlord.
- B3 The man I saw in Plaza de Bolivar is a good person (from **B1** and **B2**).

This is the new information the agent is confronted with:

- D1 The man I saw in Plaza de Bolivar is Alvaro Uribe.

...

- DN Alvaro Uribe is not a good person.

If the agent accepts **DN**, the latter conflicts with the agent present belief that **B3**.

Thus, the agent should find a way to make **B3** false. Some options:

1. Giving up **B1** or giving up **B2**.
2. Modifying **B1**, e.g., "all landlords are good people, except Alvaro Uribe."

Belief Revision: Example 2

Suppose this is (part of) what an agent believes:

B1 φ

B2 $\varphi \rightarrow \psi$

B3 ψ (from **B1** and **B2**).

This is the new information the agent is confronted with:

D1 φ'

...

DN $\neg\psi$

If the agent accepts **DN**, the latter conflicts with the agent present belief that **B3**.

Thus, the agent should find a way to make **B3** false. Some options:

1. Giving up φ or $\varphi \rightarrow \psi$.
2. Modifying φ or $\varphi \rightarrow \psi$.

Methodology

- We seek a formal theory that can model the phenomena of belief revision.
- We rely on our intuitions as to how the operation of belief revision should be performed:
  1. A belief set should be *consistent*.
  2. Any change of a belief set should *minimize loss of information*.
- However, there are many open issues. E.g.
  1. Some agents trust new information more than other agents (skeptical vs. trusting agents).
  2. There are many kind of beliefs we can include into the beliefs sets (belief about the world, modal beliefs, preferences, desires, values, expectations, etc.).
  3. ...

Remark: Notice the methodological differences between using formal methods in philosophy, formal semantics and computer science.

How to Represent Belief Sets

Two options:

1. Set of sentences (syntactic representation).
2. Set of points or worlds (semantic representation).

On the relation between the two:

- The operation '*Mod*' yields a set of point out of set of sentences:
  $Mod(\Gamma) = \{w : w \Vdash \varphi \in \Gamma\}$.
- The operation '*Th*' yields a set of sentences out of a set of points: $Th(P) = \{\varphi : w \Vdash \varphi, w \in P\}$.

Computer Science Tradition

1. Define a list of postulates (=standards of rationality) that any operation of belief revision should satisfy.
2. Define a function, procedure or algorithm that satisfies the postulates.
   This can be proven by means of *representation theorems*.

Given a belief set K (=set of formulas), and the operations of expansion $+$ and revision $*$ with formulas $\varphi \in \Lambda$, the following are the **AGM Postulates:**

P1 $K * \varphi$ is a belief set.

P2 If we revise K by φ, then $\varphi \in K * \varphi$.

P3 $K * \varphi \subseteq K + \varphi$

P4 If $\neg\varphi \notin K$, then $K + \varphi \subseteq K * \varphi$.

P5 $K * \varphi = K_{\perp}$ iff $\vdash \neg\varphi$.

P6 If $\vdash \varphi \leftrightarrow \psi$, then $K * \varphi = K * \psi$.

The Logical Tradition (from Lewis' Spheres to Belief Revision)

We can re-interpret Lewis spheres in a particular way:

- The innermost sphere represent the agent's belief set $|K|$
- The outermost spheres represent the agents doxastic disposition.
- A revision of $|K|$ by $|\varphi|$ yileds $|K * \varphi|$, where: $|K * \varphi| = |\varphi| \cap S$, where S is the smallest sphere such that $S \cap |\varphi| \neq \emptyset$.

Adam Grove proved that $|K * \varphi|$ satisfies the **AGM** postulates. This showed the connection between counterfactuality and belief revision.

A Logic for BR: Basic Ingredients

1. The **logical space** (=set of points) representing all possible states of the world (from some viewpoint).
2. A **proposition** about the world is a subset of the logical space.
3. A **theory** about the world is the intersection of some propositions.
   - A **belief set** is the intersection of propositions believed by an agent.
4. A **belief state** is more complicated than a belief set:
   - it includes not only what an agent actually believes.
   - it should include a dynamic perspective (modeling belief change).
   - e.g., it also includes how an agent would react, if confronted with new and contradictory beliefs.
   - suggestion: representing beliefs states as **hypertheories**.

A Mathematical Structure for Belief Revision

$(\wp(\mathcal{W}), \cap, \cup, -, \mathcal{W}, \emptyset)$

- A set of points or worlds $\mathcal{W}$.
- The empty set $\emptyset$.
- The set of subsets of $\mathcal{W}$, i.e., the powerset of $\mathcal{W}$: $\wp(\mathcal{W})$
- Operations on $\wp(\mathcal{W})$: intersection $\cap$, union $\cup$, substraction $-$.

- A theory T (in the semantical sense) is such that $T = \bigcap S$, with $S \subseteq \wp(\mathcal{W})$.
- A hypertheory $\mathcal{H}$ is such that $H \subseteq \wp(\mathcal{W})$, and:

  NO-E: $\mathcal{H} \neq \emptyset$;

  NEST: for any $X, Y \in \mathcal{H}$, we have $X \subseteq Y$ or $Y \subseteq X$;

  LIMA: let $C = \{X \in \mathcal{H} : X \cap P \neq \emptyset\}$, for $P \subseteq \mathcal{W}$.
  If $C \neq \emptyset$, then $\bigcap C \in C$.
  (That is: if there is a $X \in \mathcal{H}$ with $P \cap X \neq \emptyset$,
  then there is a *smallest* $X \in \mathcal{O}$ with $P \cap X \neq \emptyset$.)

The Language of Dynamic Doxastic Logic (DDL)

$\mathbf{B}\varphi$, $\mathbf{b}\varphi$
$\mathbf{K}\varphi$, $\mathbf{k}\varphi$
$[*\varphi]\chi$, $(*\varphi)\chi$

Remark: The operators **B** and **K** (and their duals) and $*$ operates only on **purely Boolean formulas**. Thus, we are only modeling belief about the world, and not modal beliefs.

The Semantics of DDL

To each purely Boolean formula φ we assign a set of points in which φ is true: $|\varphi|$. This is done recursively in the usual way.

$H, w \Vdash \varphi$ iff $w \in |\varphi|$.

$H, w \Vdash \mathbf{B}\varphi$ iff $\bigcap H \subseteq |\varphi|$

$H, w \Vdash \mathbf{K}\varphi$ iff $\bigcup H \subseteq |\varphi|$

$H, w \Vdash \mathbf{b}\varphi$ iff $\bigcap H \cap |\varphi| \neq \emptyset$

$H, w \Vdash \mathbf{k}\varphi$ iff $\bigcup H \cap |\varphi| \neq \emptyset$

$H, w \Vdash [*\varphi]\chi$ iff $H', w \Vdash \chi$, for all H' with $(H, H') \in R * \varphi$.

$H, w \Vdash (*\varphi)\chi$ iff $H', w \Vdash \chi$, for some H' with $(H, H') \in R * \varphi$

$(H, H') \in R * \varphi$ iff $\bigcap H' = \bigcap C$, where $C = \{X \in H \colon X \cap |\varphi| \neq \emptyset\}$.

Some Axioms of DDL

$\chi \leftrightarrow [*\varphi]\chi$, if χ is purely Boolean.

If $\vdash \varphi \leftrightarrow \psi$, then $[*\varphi]\chi \leftrightarrow [*\psi]\chi$ (like **P6**).

$\mathbf{b}\varphi \rightarrow ([*\varphi]\mathbf{B}\chi \leftrightarrow \mathbf{B}(\varphi \rightarrow \psi))$ (like **P4**)

$[*\varphi]\mathbf{K}\varphi$ (like **P2**)

Modus Ponens Fails

Assume

A1 $\mathbf{B}\pi$

A2 $[*\pi][\neg\rho]\mathbf{B}\alpha$

It does not follow that

C $[\neg\rho]\mathbf{B}\alpha$

Iteration

Problem How do we built a complete new hypertheory H' out of the old one H?

A Related Area

Problem Belief merging: what happens when two hypertheory merge?