An interactive demonstration of counterfactual truth conditions

Bachelor Thesis

Andreas Paul Bruno Lönne
loenne@campus.tu-berlin.de

Technische Universität Berlin
discourse Degree program: Bachelor Informatik / Computer Science

Abstract

In this thesis, I address the scarcity of online resources, showcasing counterfactual truth conditions in an intuitive and digestible manner. To this end I (i) formulate a semantic satisfiability game for counterfactual sentences; (ii) prove its correctness; (iii) prove that it always halts after a finite number of moves; (iv) develop a browserbased web-application, that makes the semantic game of counterfactuals playable.

1 Introduction

In this thesis i make the attempt to create an application that is able to convey Lewis' counterfactual truthconditions by way of a semantic two-player game and inspire players to learn more about counterfactuals.
Background and literature
To this end i defined a semantic game of counterfactuals and implemented a browserbased demonstration game.
The document is laid out as follows:
First i will explain counterfactuals.
First i will begin by introducing the counterfactual logic im basing the semantic game of counterfactuals on. Then i will give game-theoretical definitions and formulate two versions of the semantic game. After that ...

2 Counterfactuals

Counterfactuals are statements about what might or would have been the case, if things took place differently than they did. One may think to themselves "If I had not forgotten about my appointment, I would not have been late". Or one may wonder "If Alexander the Great had not died at the age of 32 and attacked europe, would the Romans have defeated him?". Such *counterfactual thought*—that is the thought of alternate outcomes—is essential for reasoning, deduction and cognitive function. [Byr16] Due to its abundance in human thought, the ability to imagine alternate realities seems trivial to most. But making rigorous statements or claims about them is difficult. This is because communicating a complete

and consistent account of the state of affairs of an alternate reality—similar in complexity to ours— is difficult, if not impossible. One may attempt to circumvent this issue by giving the state of affairs of an alternate reality as a deviation from the state of affairs of reality. But consider this. Take our previous example about Alexander the Great and assume that the imagined alternative reality is identical to our reality, except that Alexander the Great did not die at the age of 32 and attacked europe. Now suppose we know about reality that Alexanders troops remained outside of europe. Then this should also be the case in the alternate reality, we attempt to describe. If we are to assume that an army cannot be in two places at the same time and Alexander could not have attacked europe without his army, then Alexanders troops could not have remained outside of europe and attacked europe at the same time. We find, that simply deviating from our own reality in a few concrete ways may produce internally inconsistent alternate realities, which cannot be alternate realities, because they could never exist in the first place.
Instead of trying to fully describe an alternate reality, we want to assert some statement over,

- possible world is complete and consistent account of alternate reality - we can compare possible worlds with respect to their similarity - counterfactual structure antecedent -> consequent - interpret counterfactual to mean something like "In all the closest alternate realities, where antecedent, then consequent" *- disagreement on implicit premises*
- incongruent premises
- possible world semantics (there is a real world, a world is a way the real world could have been (semantic trick or truth?))
Counterfactual constructions mitigate this issue by assuming the alternate reality to be as similar to reality as possible and deviate in the states of affairs, that are explicitly mentioned to differ.
—> possible worlds intro (closest world, where antecedent holds)
// TODO: identical not possible
- implicit antecedents (hard to define, disagreement about them)
- invoke possible worlds
- accessiblity (worlds too remote)
- similarity
- explain sphere of accessibility

3 The semantic game of counterfactuals

two-player game about proving counterfactual sentences
I will give two formulations
- the first making the limit assumption
- the second not doing so

3.1 Counterfactual logic

This section is concerned with defining counterfactual logic. First it provides a brief definition of well-formed counterfactual formulas, then introduces possible world semantics and concludes by defining the truth conditions of counterfactual logic.

3.1.1 Counterfactual formulas

We call $\Phi = \{\varphi, \psi, ...\}$ the set of all well-formed counterfactual formulas.

Definition 1 (Well-formed counterfactual formula)**.** Given an infinite set of atomic formula symbols $Atoms = \{x, y, ...\}$ and an alphabet $A = \{\bot, \top, \neg, \vee, \wedge, \Diamond, \Box, \Diamond\!\!\rightarrow, \Box\!\!\rightarrow\} \cup Atoms$, the structure of every well-formed counterfactual formula is expressed through the following Backus-Naur form.

$$\varphi, \psi ::= \bot \mid x \mid \neg\varphi \mid \Box\varphi \mid \Diamond\varphi \mid \varphi \vee \psi \mid \varphi \wedge \psi \mid \varphi \mathrel{\Box\!\!\rightarrow} \psi \mid \varphi \mathrel{\Diamond\!\!\rightarrow} \psi \tag{1}$$

3.1.2 Possible worlds semantics

In order to evaluate the non-truth-functional connectives of counterfactual logic, we define a variation of the Kripke structure in [Kri63]. To evaluate Lewis' counterfactual truth conditions [Lew73], we also introduce a notion of similarity between possible worlds to the accessibility relation.

Definition 2 (counterfactual kripke structure)**.** The *counterfactual kripke structure* is an ordered triple $(W, \rightsquigarrow, F)$, where $W = \{w, v, ...\}$ is the set of all "possible worlds", $\rightsquigarrow : W \times \mathbb{R} \times W$ is the similarity relation, and $F\colon W \to 2^{Atoms}$ is an assignment of each world, to a set of atomic formulas.

Let us explain further the members of our counterfactual kripke structure.
The set of all possible worlds W represents all complete and self-consistent ways reality could have been. F in turn, describes the state of affairs at each world. It assigns to each world a set of atomic propositions, that are the case at it. The similarity relations purpose is twofold. It serves as an accessibility relation, restricting accessibility between worlds and also carries information about comparative similarity between worlds. In particular this means for any two worlds w and v, that iff $w \overset{r}{\rightsquigarrow} v$, then v is called accessible from w. $w \overset{r}{\rightsquigarrow} v$ also means, that r is a measure of similarity between w and v, from the standpoint of w. We can suppose another world v', with $w \overset{r'}{\rightsquigarrow} v'$ and compare r and r'. If $r < r'$, we can conclude, that v is more similar to w than v', from the standpoint of w. Additionally we note, that for our purposes $\rightsquigarrow$ will always contain a tuple $(w, 0, w)$ for each world $w \in W$. This is because we assume analogously to Kripke, that worlds are self-similar and are accessible from themselves. Finally we introduce $W_w = \{w' \mid w \overset{r}{\rightsquigarrow} w'\}$ as a shorthand for the set of all accessible worlds, from a world w.

3.1.3 Truth conditions of counterfactual logic

Given a counterfactual kripke structure $S = (W, \rightsquigarrow, F)$ and a world w,
$S, w \nvDash \bot$.
$S, w \vDash \top$.
$S, w \vDash x$ iff $x \in F(w)$.
$S, w \vDash \neg\varphi$ iff $S, w \nvDash \varphi$.
$S, w \vDash \varphi \vee \psi$ iff ($S, w \vDash \varphi$ or $S, w \vDash \psi$).
$S, w \vDash \varphi \wedge \psi$ iff ($S, w \vDash \varphi$ and $S, w \vDash \psi$).
$S, w \vDash \Diamond\varphi$ iff a world $w' \in W_w$ exists, such that $S, w' \vDash \varphi$.
$S, w \vDash \Box\varphi$ iff for every world $w' \in W_w$, it is true that $S, w' \vDash \varphi$.
$S, w \vDash \varphi \mathrel{\Diamond\!\!\rightarrow} \psi$, iff both

(1) a world $w' \in W_w$ exists, such that $S, w' \vDash \varphi$,

(2) for every world w'', for which an r'' exists, such that $w \overset{r''}{\rightsquigarrow} w''$ and $S, w'' \vDash \varphi \wedge \neg\psi$ are true, a world w^* and an r^* exist, such that $r^* \leq r''$ and $w \overset{r^*}{\rightsquigarrow} w^*$ and $S, w^* \vDash \varphi \wedge \psi$.

$S, w \vDash \varphi \boxdot\!\!\rightarrow \psi$, iff either

(1) no world $w' \in W_w$ exists, such that $S, w' \vDash \varphi$.

(2) a world w' and an r exist, such that $w \overset{r}{\rightsquigarrow} w'$ and $S, w' \vDash \varphi$ and for each world w^*, for which an r^* exists, such that $r^* \leq r$ and $w \overset{r^*}{\rightsquigarrow} w^*$, it is true that $S, w^* \vDash \psi \vee \neg\varphi$.

3.2 Semantic game

This section offers an overview and basic definitions for the semantic game of counterfactuals. The semantic game of counterfactuals is a sequential two-player satisfiability game, wherein a *defender d* tries to prove a counterfactual formula and an *attacker a* tries to disprove it. The game is played on a labeled transition system, where the respective active player of a game state chooses a transition to another game state.

Definition 3 (Game state)**.** A game state is a tuple that can take any one of the forms

(1) $(\varphi, w)_p$

(2) $(\varphi, w, e)_p$

(3) $(\varphi, w, w', r)_p$

where $\varphi \in \Phi$; $w, w' \in W$; $r \in \mathbb{R}$ and $p, e \in \{a, d\}$.

Of these types of game states (1) is the most common. It carries a counterfactual formula, a current world and is indexed with the active player. It is employed to resolve atomic formulas $x, y, ...$ and the propositional symbols $\top, \neg, \vee, \Diamond$; while also serving as an initial and final state type for other resolutions. (2) is a type of game state, that represents a player temporarily becoming active and choosing a state transition. e holds the previous active player, which will resume being active after the next state transition. Lastly state type (3) is an extension of (1), describing a game state where a sphere of accessibility has been chosen. w' is a sphere-delimiting world, lying exactly on the surface of said sphere of accessibility with radius r. We will look at this in further detail in the next subsetion.
We call the set of all game states G.

Definition 4 (Set of Labels)**.** Σ is a finite set of labels. Σ = {

- Top,
- Negation,
-

}

Transitions of the labeled transition system are called moves.

Definition 5 (Move)**.** A move is a transition function m : *gamestate* $\rightarrow$ *gamestate*

(1)

(2)

(3)

Definition 6 (Labeled transition system)**.** A labeled transition system is a tuple $(G, \Sigma, \rightarrow)$

Definition 7 (Semantic game of counterfactuals)**.** A labeled transition system is a tuple $(G, \Sigma, \rightarrow)$

3.3 Formulation of the semantic games rules

- Rules
(stuck player loses) - Employment of the Limit assumption - alternative semantic game rules

3.4 Game in game-theoretical formulation

This section is concerned with explaining core ideas and definitions related to semantic two-player games. - game is like a conversation between players - show something through their interplay - show truth of counterfactual formula

Definition 8.

- game states (formula, current world, player)
- game outcomes (winning/losing)
- moves
- play (sequence of moves)
- winning play
- game tuple (state, moves)
- strategy (mapping from game states to moves)

3.5 Correctness proof

3.6 Termination proof

4 The educational game

4.1 Correspondance between semantic game and computer game

5 Results

5.1 Player feedback

References

[Byr16] Ruth MJ Byrne. Counterfactual thought. *Annual review of psychology*, 67(1):135–157, 2016.

[Kri63] Saul A. Kripke. Semantical considerations on modal logic. *Acta Philosophica Fennica*, 16:83–94, 1963.

[Lew73] David K. Lewis. *Counterfactuals*. Basil Blackwell, Oxford, 1973.