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 truth conditions 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 I am 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 have been punctual". 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 alternate 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 attempted 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 are not ways the world could have been.
To avoid this problem we forgo describing alternate realities alltogether. We call a complete and consistent way the world is or could have been, a possible world. And agree that we refer to a possible world, most similar to ours, where our stipulation is true. So our example is to be read as "In a world most similar to our own, where Alexander the Great did not die at the age of 32 and attacked europe, the romans would have defeated him". In this way—although we may not know the state of affairs of a possible world—we are able to assign definite truth values to our counterfactual sentence for any possible state of affairs at that world. However we need to note, that this approach assumes the notion of comparative similarity between possible worlds. Which means that given any 3 worlds w, v_1, v_2, with respect to their overall similarity from the standpoint of w, either

- w is more similar to v_1, than to v_2,
- w is more similar to v_2, than to v_1,
- or w is equally similar to v_1 and v_2.

While the notion of an aggregate overall similarity between possible worlds appears justified at first glance, it has been subject of contention. [Mor10]

2.1 Lewis' counterfactual operators

With these introductory thoughts out of the way, let us talk in greater detail about the counterfactual operators themselves. Lewis introduces the counterfactual would $\boxright$ and counterfactual might $\diamondright$ operators as binary modal operators. [Lew73] When we write $\varphi \boxright \psi$, we call the formula φ the antecedent and the formula ψ the consequent. We may informally rewrite one of our former examples as "I did not forget about my appointment $\boxright$ I was punctual" and read it the following way.
Read $\varphi \boxright \psi$ as "If it were the case that φ, then it would be the case that ψ", and
read $\varphi \diamondright \psi$ as "If it were the case that φ, then it might be the case that ψ".
Lewis defines the truth conditions of his operators with respect to a system of spheres, that is defined as follows.

Let $ be an assignment to each possible world i of a set $\$_i$ of sets of possible worlds. Then $ is called a (centered) system of spheres, and the members of each $\$_i$ are called spheres around i, if and only if, for each world i, the following conditions hold.

(C) $\$_i$ is centered on i; that is, the set $\{i\}$ having i as its only member belongs to $\$_i$.

(1) $\$_i$ is nested; that is, whenever S and T belong to $\$_i$, either S is included in T or T is included in S.

(2) $\$_i$ is closed under unions; that is, whenever $\mathcal{S}$ is a subset of $\$_i$ and $\bigcup \mathcal{S}$ is the set of all worlds j such that j belongs to some member of $\mathcal{S}$, $\bigcup \mathcal{S}$ belongs to $\$_i$.

(3) $\$_i$ is closed under (nonempty) intersections; that is, whenever $\mathcal{S}$ is a nonempty subset of $\$_i$ and $\bigcap \mathcal{S}$ is the set of all worlds j such that j belongs to every member of $\mathcal{S}$, $\bigcap \mathcal{S}$ belongs to $\$_i$.

Fig. 1: Lewis' (centered) system of spheres

Illustrate example system of spheres and explain $

A world where φ holds is called a φ-world. And his counterfactual truth conditions are:

$\varphi \diamond\!\!\rightarrow \psi$ is true at a world i (according to a system of spheres $) if and only if both

(1) some φ-world belongs to some sphere S in $\$_i$, and

(2) every sphere S in $\$_i$ that contains at least one φ-world contains at least one world where $\varphi \wedge \psi$ holds.

Fig. 2: Lewis' counterfactual might truth conditions

$\varphi \mathrel{\Box\!\!\rightarrow} \psi$ is true at a world i (according to a system of spheres $) if and only if either

(1) no φ-world belongs to any sphere S in $\$_i$, or

(2) some sphere S in $\$_i$ does contain at least one φ-world, and $\varphi \rightarrow \psi$ holds at every world in S.

Fig. 3: Lewis' counterfactual would truth conditions

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.

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

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)**.** A *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 $F : W \to 2^{Atoms}$ 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 $S, w' \vDash \varphi$.
$S, w \vDash \varphi \diamondsuit\!\!\rightarrow \psi$, iff both

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

(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 $S, w^* \vDash \varphi \wedge \psi$.

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

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

(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 The semantic game of counterfactuals

This subsection offers an overview and introduction to 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 defender begins the game as the *active player*, with a counterfactual formula at a possible world. As the active player makes moves, the counterfactual formula is resolved step by step, until a player resolves a $\top$ and wins or cannot make any move and loses. Throughout the game, moves may also change the active player and possible world.

3.2.1 Game state

Game states form the backbone of the semantic game of counterfactuals, by describing its game positions. Each game state describes the game at a discrete point in time, inbetween moves.

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, the 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 game 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.

3.2.2 Moves

When the active player transforms the current game state into another and progresses the game, we call that a move. Formally we represent this in the following way.

Definition 4 (Move)**.** A move m is a tuple $m \in G \times G$.

The first element of the tuple is the initial game state, while the second one is the resulting game state. We define the set of moves $\rightarrow$ specific to the semantic game of counterfactuals.

Definition 5 (Moves)**.** $\rightarrow$ is a set of moves, such that for each world w, player p ..., the following is true. And for each p, $np \in \{a, d\} \setminus \{p\}$.

$$(x, w)_p \xrightarrow{[x \in F(w)]} (\top, w)_p \tag{2}$$

$$(\neg\varphi, w)_p \to (\varphi, w)_{np} \tag{3}$$

$$(\varphi \vee \psi, w)_p \to (\varphi, w)_p \tag{4}$$

$$(\varphi \vee \psi, w)_p \to (\psi, w)_p \tag{5}$$

$$(\varphi \wedge \psi, w)_p \to (\varphi \wedge \psi, w, p)_{np} \tag{6}$$

$$(\varphi \wedge \psi, w, e)_p \to (\varphi, w)_e \tag{7}$$

$$(\varphi \wedge \psi, w, e)_p \to (\psi, w)_e \tag{8}$$

$$(\Diamond\varphi, w)_p \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi, w')_p \tag{9}$$

$$(\Box\varphi, w)_p \to (\Box\varphi, w, p)_{np} \tag{10}$$

$$(\Box\varphi, w, e)_p \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi, w')_e \tag{11}$$

$$(\varphi \Diamond\!\!\rightarrow \psi, w)_p \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi \Diamond\!\!\rightarrow \psi, w, w', r)_{np} \tag{12}$$

$$(\varphi \Diamond\!\!\rightarrow \psi, w, w', r)_p \to (\varphi \wedge \psi, w')_{np} \tag{13}$$

$$(\varphi \Diamond\!\!\rightarrow \psi, w, w', r)_p \xrightarrow{[w \overset{r^*}{\rightsquigarrow} w^*, r^* < r]} (\neg\varphi, w^*)_{np} \tag{14}$$

$$(\varphi \Box\!\!\rightarrow \psi, w)_p \to (\varphi \Box\!\!\rightarrow \psi, w, p)_{np} \tag{15}$$

$$(\varphi \Box\!\!\rightarrow \psi, w, e)_p \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi \Box\!\!\rightarrow \psi, w, w', r)_e \tag{16}$$

$$(\varphi \Box\!\!\rightarrow \psi, w, w', r)_p \to (\neg\varphi \vee \psi, w')_p \tag{17}$$

$$(\varphi \Box\!\!\rightarrow \psi, w, w', r)_p \xrightarrow{[w \overset{r^*}{\rightsquigarrow} w^*, r^* < r]} (\varphi, w^*)_p \tag{18}$$

Fig. 4: Moves of the semantic game of counterfactuals

3.2.3 The transition system

The moves, that can be played in a given game state, are determined by a transition system L. It contains transitions between game states, which we call moves. To arrive at our definiton for L, let us first define game states and transitions themselves.

Definition 6 (Transition system)**.** A transition system is a tuple $(G, \to)$, where G is the set of all game states and $\to$ is a transition relation.

3.3 Limit Assumption

- explain limit assumption
- justify limit assumption (finite graph satisfies it anyways, nicer interactive presentation)
- alternative semantic game rules

3.4 Alternative game rules

$$(x, w)_p \xrightarrow{[x \in F(w)]} (\top, w)_p \tag{19}$$

$$(\neg\varphi, w)_p \rightarrow (\varphi, w)_{np} \tag{20}$$

$$(\varphi \vee \psi, w)_p \rightarrow (\varphi, w)_p \tag{21}$$

$$(\varphi \vee \psi, w)_p \rightarrow (\psi, w)_p \tag{22}$$

$$(\varphi \wedge \psi, w)_p \rightarrow (\varphi \wedge \psi, w, p)_{np} \tag{23}$$

$$(\varphi \wedge \psi, w, e)_p \rightarrow (\varphi, w)_e \tag{24}$$

$$(\varphi \wedge \psi, w, e)_p \rightarrow (\psi, w)_e \tag{25}$$

$$(\Diamond\varphi, w)_p \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi, w')_p \tag{26}$$

$$(\Box\varphi, w)_p \rightarrow (\Box\varphi, w, p)_{np} \tag{27}$$

$$(\Box\varphi, w, e)_p \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi, w')_e \tag{28}$$

//TODO: Check truth conditions

$$(\varphi \diamond\!\!\rightarrow \psi, w)_p \rightarrow (\varphi \diamond\!\!\rightarrow \psi, w, p)_{np} \tag{29}$$

$$(\varphi \diamond\!\!\rightarrow \psi, w, e)_p \rightarrow (vac, \varphi \diamond\!\!\rightarrow \psi, w)_e \tag{30}$$

$$(vac, \varphi \diamond\!\!\rightarrow \psi, w)_p \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi, w')_{np} \tag{31}$$

$$(\varphi \diamond\!\!\rightarrow \psi, w, e)_p \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi \diamond\!\!\rightarrow \psi, w, w', r)_e \tag{32}$$

$$(\varphi \diamond\!\!\rightarrow \psi, w, w', r)_p \rightarrow (\neg\varphi \vee \psi, w')_p \tag{33}$$

$$(\varphi \diamond\!\!\rightarrow \psi, w, w', r)_p \xrightarrow{[w \overset{r^*}{\rightsquigarrow} w^*, r^* \leq r]} (\varphi \wedge \psi, w^*)_p \tag{34}$$

//TODO: Accurate truth conditions

$$(\varphi \,\Box\!\!\rightarrow \psi, w)_p \rightarrow (\Box\neg\varphi, w)_p \tag{35}$$

$$(\varphi \,\Box\!\!\rightarrow \psi, w)_p \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi \,\Box\!\!\rightarrow \psi, w, w', r)_{np} \tag{36}$$

$$(\varphi \,\Box\!\!\rightarrow \psi, w, w', r)_p \rightarrow (\varphi, w')_{np} \tag{37}$$

$$(\varphi \,\Box\!\!\rightarrow \psi, w, w', r)_p \xrightarrow{[w \overset{r^*}{\rightsquigarrow} w^*, r^* \leq r]} (\neg\varphi \vee \psi, w^*)_{np} \tag{38}$$

Fig. 5: alternate rules of the semantic game of counterfactuals

3.5 Playing the game

Definition 7 (Semantic game of counterfactuals)**.** The semantic game of counterfactuals is an ordered tuple (I, E, L), where $I \in G$ is the initial game state; $E \subseteq G$ is the set of ending game states; L is a transition system for game states.

TODO: Add ending game states - Rules
(stuck player loses)
- game is like a conversation between players
- show something through their interplay
- show truth of counterfactual formula
- Game produces a play
- game states (formula, current world, player)
- game outcomes (winning/losing)
- moves
- play (sequence of moves)
- winning play
- game tuple (state, moves)
- subgame
- strategy (mapping from game states to moves (positional game, thus state-based is okay, otherwise partial mapping from partial plays to moves))

3.6 Correctness proof

3.7 Termination proof

4 The educational game

4.1 Correspondance between semantic game and computer game

4.2 Artificial intelligence

Algorithm 1 Adjusted MiniMax Algorithm

```
Input: s (game state)
  function MiniMax(s)                                  ▷ Calculate best move
    Q ← [s]                                  ▷ Setup queue, expanded node and score lists
    E ← []
    SC ← []
    while Q not empty do
      current ← POP(Q)
      e ← false
      states ← NEXTSTATES(s)
      Let n
      if c not in E then
        PUSH(E, current)
        e ← true
        for each state in states do
          if state not in E then
            if state in Q then
              REMOVEEACH(Q, state)
            end if
            PUSH(Q, state)
          end if
        end for
      end if
      if e = false then // TODO:
      end if
    end while
  end function
  NEXT calculates the next game states and returns them
```

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.

[Mor10] Michael Morreau. It simply does not add up: Trouble with overall similarity. *The journal of philosophy*, 107(9):469–490, 2010.