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 Lewis' counterfactual truth conditions in an intuitive and digestible manner. To this end, I (i) formulate a semantic reachability game for counterfactual sentences; (ii) prove its correctness; (iii) prove that it always halts after a finite number of moves; (iv) develop a browser-based 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
- contentious subject within philosophy (Lewis realism about possible worlds (Are possible worlds real?) and comparative similarity (Can you weigh differences up against each other?)) - academic standstill for about 50 years - you can find this demo under ... on github To this end i defined a semantic game of counterfactuals and implemented a browser-based 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 all-together. 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]

3 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 $\Box\!\!\rightarrow$ and counterfactual might $\Diamond\!\!\rightarrow$ operators as binary modal operators. [Lew73] When we write $\varphi \ \Box\!\!\rightarrow \ \psi$, 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 $\Box\!\!\rightarrow$ I was punctual" and read it the following way. Read $\varphi \ \Box\!\!\rightarrow \ \psi$ as "If it were the case that φ, then it would be the case that ψ", and read $\varphi \ \Diamond\!\!\rightarrow \ \psi$ as "If it were the case that φ, then it might be the case that ψ".

3.1 Lewis' system of spheres

The truth conditions of the counterfactual operators are stated with respect to a structure, called a *system of spheres*. Lewis defines a system of spheres 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' system of spheres

Now let us first take a broad perspective upon this definition. Most generally a system of spheres is an assignment of a set of sets of possible worlds to each world. The assigned sets exhibit the four characteristics (C), (1), (2) and (3); and are called sets of spheres. In totality such a system of spheres serves to codify the relationships of accessibility and comparative similarity between possible worlds. To this end each set of spheres describes an ordering of possible worlds with respect to similarity to the world it is assigned to. Meaning, that the set of spheres $\$_i$ describes the comparative similarity to- and accessibility from i for each world. This is briefly illustrated by the next example.

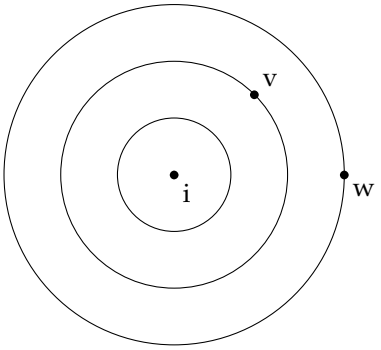

Fig. 2: A set of spheres $\$_i = \{\{i\}, \{i, v\}, \{i, v, w\}\}$

In this example we look at the worlds i, v and w; and consider the set of spheres $\$_i$. The spheres themselves are sets of possible worlds and are visualized through circles. A world inside or on the circumference of a circle is contained within the corresponding sphere. Crucially, the meaning of a sphere around i is that each world contained in it is more similar to i than each world not contained in it. For our example this means that i is more similar to itself than any other world; i and v are more similar to i than any other world; and that i, v and w are more similar to i than any other world. Regarding accessibility, a world is said to be accessible from i, if it is contained in any sphere around i. So we can see that all i, v and w are accessible from i. With this in mind, we can motivate the properties (C), (1), (2) and (3).

(C) It seems reasonable to assume that each world is most similar to itself and that there cannot be another world equally or more similar to it than itself. If that is the case, every sphere around a world i needs to contain the set $\{i\}$.

(1) Intuitively it may also make sense, that for each pair of spheres around the same world one should include the other. If we entertain the notion, that the spheres around a world are not nested, then one of those spheres cannot be a sphere. First, remember that a sphere around a world i is a set of possible worlds, such that each world contained in it is more similar to i, than every world not contained in it. Then suppose the set of spheres $\$_i$ is not nested. In that case two worlds v and w and the spheres $S, T \in \$_i$ exist, such that $v \in S$, $w \notin S$, $v \notin T$ and $w \in T$. Through S we know that that v is more similar to i than w. And through T we know that that w is more similar to i than v. Which cannot both be the case at the same time.

(2) Suppose that for the union $\bigcup \mathcal{S}$ of a set of spheres $\mathcal{S}$ around i, there are two worlds v, w such that $v \in \bigcup \mathcal{S}$ and $w \notin \bigcup \mathcal{S}$. Then that means, that v is, and w is not, contained in some sphere in $\mathcal{S}$. Hence v is more similar to i than w. Therefore $\bigcup \mathcal{S}$ is a set such that any world contained within it is more similar to i than any world not contained in it. We call such a set a sphere around i.

(3) Suppose that for the intersection $\bigcap \mathcal{S}$ of a nonempty set of spheres $\mathcal{S}$ around i, there are two worlds v, w such that $v \in \bigcap \mathcal{S}$ and $w \notin \bigcap \mathcal{S}$. Then that means, that v is, and w is not, contained in some sphere in $\mathcal{S}$. Hence v is more similar to i than w. Therefore $\bigcap \mathcal{S}$ is a set such that any world contained within it is more similar to i than any world not contained in it. We call such a set a sphere around i.

3.2 Lewis' counterfactual truth conditions

Then let us take a look at Lewis' counterfactual truth conditions. Take the counterfactual operators as part of some logic, where formulas like φ and ψ are evaluated with respect to some world and a system of spheres. Lewis abbreviates a world, where the formula φ holds as a φ-world. The truth conditions of Lewis' counterfactual would operator are these.

$\varphi \mathrel{\square\!\!\rightarrow} \psi$ 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

The counterfactual would evaluates to true iff either (1) or (2) is true.

(1) We describe the first case as vacuous truth. Intuit this case the same way a material implication is true, when its antecedent is false. From wrong antecedents arbitrary consequents may be inferred. Keep in mind however, that the vacuous truth of the counterfactual would is much stricter. While the material conditional only requires the antecedent to be false at the world it is evaluated at, the counterfactual would requires the antecedent to be false at every accessible world.

(2) The non-vacuous case requires the existence of a sphere throughout which $\varphi \rightarrow \psi$ holds and that that sphere also contains a φ-world. The existence of such a sphere can

roughly be understood as the existence of some accessibility restriction under which the strict conditional $\Box(\varphi \rightarrow \psi)$ comes out as non-vacuously true. Meaning that $\varphi \rightarrow \psi$ holds throughout all accessible worlds and that at least one accessible world is a φ-world. While the strict conditional is true, only if it is true in the broadest sense—meaning it is true at every accessible world—the counterfactual would conditional merely needs to be true in one of many senses. That is why Lewis' counterfactual operators are called variably strict conditionals. Each of these many senses are degrees of similarity to a world, that are described by spheres around that world. To motivate why overall similarity between worlds is used as an accessibility restriction consider this example.

"If Kennedy had pressed the red button, there would have been nuclear war" (1)

We consider this to be true. Without much trouble, one can make up many counterexamples to this counterfactual. Imagine for a moment the possibility that Kennedy had pressed the red button and the button malfunctioned. In that case Kennedy pressed the button, but there is no nuclear war. Obviously that is not meant. When we say something like "If it were the case that A, then it would be the case that B", we implicitly mean "If it were the case that A and things were pretty much as they are otherwise, then it would be the case that B". Our rebuttal therefore may be that there is at least one possible world where Kennedy pressed the red button and the red button did not malfunction that is more similar to the actual world than any possible world where Kennedy pressed the red button and the button did malfunction. If this is the case, we may claim that our example counterfactual is true, although there are certain senses in which it is not true. This is why a counterfactual would operator is true, if there is at least one sense, represented through a sphere of accessibility, in which it is true. Regarding the system of spheres it means that a counterfactual would operator is non-vacuously true, unless each sphere around the world it is evaluated at contains a refuting world that serves as a counterexample or no sphere around the world it is evaluated at contains a φ-world.
The *counterfactual might* is defined analogously.

$\varphi \Diamond\!\!\rightarrow \psi$ 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. 4: Lewis' counterfactual might truth conditions

Note the differences in quantification and truth conditions. The counterfactual would's existential quantifier for the existance of a sphere is replaced by a universal quantifier requiring every sphere containing a φ-world to also contain a $(\varphi \wedge \psi)$-world. While the counterfactual would requires each world of a sphere to fulfill its criteria, the counterfactual might requires one world of each sphere, that contains a φ-world, to fulfill its criteria. Notice also that the counterfactual might does not have two separate truth cases and is true in only one case. But since the counterfactual would and counterfactual might are interdefinable, we wont go into further detail here.

$$\varphi \Box\!\!\rightarrow \psi = \neg(\varphi \Diamond\!\!\rightarrow \neg\psi) \tag{2}$$

$$\varphi \Diamond\!\!\rightarrow \psi = \neg(\varphi \Box\!\!\rightarrow \neg\psi) \tag{3}$$

4 Counterfactual logic

This section is concerned with defining counterfactual logic. First it provides a brief definition of well-formed counterfactual formulas, then introduces the structure they are evaluated on and concludes by defining the truth conditions of counterfactual logic.

4.1 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 $A = \{\bot, \top, \neg, \vee, \wedge, \Diamond, \Box, \boxright, \Diamond\!\!\!\!-\!\!\!\!\!\rightarrow\} \cup Atoms$, the structure of every well-formed counterfactual formula is expressed through the following Backus-Naur form.

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

4.2 Counterfactual kripke structure

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 $F\colon W \to 2^{Atoms}$ 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.

4.3 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 $S, w' \vDash \varphi$.
$S, w \vDash \varphi \boxrightarrow \psi$, 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 $S, w^* \vDash \neg\varphi \vee \psi$.

$S, w \vDash \varphi \Diamond\!\!\rightarrow \psi$, 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 $S, w^* \vDash \varphi \wedge \psi$.

5 The semantic game of counterfactuals

This section offers an overview and introduction to the semantic game of counterfactuals. The semantic game of counterfactuals is a sequential two-player reachability 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. The active player is a role assigned to the player that can currently make moves and whose typical objective it is to prove the current formula through play. As the active player makes moves, the counterfactual formula is resolved progressively 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 current world.

5.1 Definition of the semantic game

The game is akin to a discussion where players make arguments and refute them. Weak arguments do not hold up to scrutiny, but strong ones do. The players' adversarial interplay allows the discovery and dismissal of arguments, such that the reason for truth or falsity of a claim is revealed. Of course, this is only possible when the players make the best possible moves. We shall call such play optimal play. We describe the game the following way.

Definition 3 (Semantic game of counterfactuals)**.** The semantic game of counterfactuals $\Im[\varphi, w] = (I, E, L, S)$ to prove the formula φ at the world w is an ordered tuple, where $I = (\varphi, w)_d$ is the initial game state; $E = \{(\top, w)_p \mid w \in W, p \in \{a, d\}\}$ is the set of winning game states; L is a transition system for game states; and $S = (W, \rightsquigarrow, F)$ is a counterfactual kripke structure.

G is the set of all game states. We will describe the structure of its elements in the upcoming subsections. The initial game state I varies from game to game. It contains the formula that ought to be proven and the world it ought to be proven at. It is always indexed by the initial active player, that is the defender. The set of winning game states contain all game states in which a player wins. In each winning game state it is the currently active

player that wins and the respective other player that loses. This set includes all game states where the formula is resolved to a $\top$-symbol. It is also possible for a player to win without reaching a winning game state. This is the case when the active player is unable to make a move and in other words stuck. When this happens, the active player loses and the active player's opponent i.e. the inactive player wins. Next we will define the transition system L.

5.2 The transition system

The moves that can be played in a given game state are determined by a transition system L.

Definition 4 (Transition system)**.** A transition system is a tuple (G, R), where G is the set of all game states and $R : G \times G$ is a transition relation.

It contains the set of all game states G and the set R of transitions between game states which we call moves. In later sections we will define two such sets of moves. One general formulation $\rightarrow$ and one simplified formulation $\rightarrow_l$ employing the limit assumption. One may take either as R. In order to arrive at our definition for $\rightarrow$ and $\rightarrow_l$, let us first define game states and transitions.

5.3 Game states

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 in-between moves. We call the set of all game states G.

Definition 5 (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 subsection.

5.4 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 through a tuple of game states.

Definition 6 (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 7 (Moves)**.** $\rightarrow$ is a set of moves, such that it contains exactly those tuples that fit any of the forms described by the following 19 types of moves. For each of those types of moves variables are quantified separately. Additionally, square brackets above the relation symbol are used to annotate restrictions imposed upon the quantification of those variables. In case of (1) for example x is restricted to range across atomic formulas contained in $F(w)$. Furthermore, it is always true that $\varphi, \psi \in \Phi$ are counterfactual formulas; $x \in Atoms$ is an atomic formula; $w, w', w^* \in W$ are worlds; $p, e \in \{a, d\}$ are players; $np \in \{a, d\} \setminus \{p\}$ is the opponent of the player p; and $r, r^* \in \mathbb{R}$.

$$(x, w)_p \xrightarrow{[x \in F(w)]} (\top, w)_p \tag{1}$$

$$(\neg\varphi, w)_p \rightarrow (\varphi, w)_{np} \tag{2}$$

$$(\varphi \vee \psi, w)_p \rightarrow (\varphi, w)_p \tag{3}$$

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

$$(\varphi \wedge \psi, w)_p \rightarrow (\varphi \wedge \psi, w, p)_{np} \tag{5}$$

$$(\varphi \wedge \psi, w, e)_p \rightarrow (\varphi, w)_e \tag{6}$$

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

$$(\Diamond\varphi, w)_p \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi, w')_p \tag{8}$$

$$(\Box\varphi, w)_p \rightarrow (\Box\varphi, w, p)_{np} \tag{9}$$

$$(\Box\varphi, w, e)_p \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi, w')_e \tag{10}$$

$$(\varphi \; \Box\!\!\rightarrow \psi, w)_p \rightarrow (\Box\neg\varphi, w)_p \tag{11}$$

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

$$(\varphi \; \Box\!\!\rightarrow \psi, w, w', r)_p \rightarrow (\varphi, w')_{np} \tag{13}$$

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

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

$$(\varphi \; \Diamond\!\!\rightarrow \psi, w, e)_p \rightarrow (\Box\neg\varphi, w)_p \tag{16}$$

$$(\varphi \; \Diamond\!\!\rightarrow \psi, w, e)_p \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi \; \Diamond\!\!\rightarrow \psi, w, w', r)_e \tag{17}$$

$$(\varphi \; \Diamond\!\!\rightarrow \psi, w, w', r)_p \rightarrow (\varphi, w')_{np} \tag{18}$$

$$(\varphi \; \Diamond\!\!\rightarrow \psi, w, w', r)_p \xrightarrow{[w \overset{r^*}{\rightsquigarrow} w^*, r^* \leq r]} (\varphi \wedge \psi, w^*)_p \tag{19}$$

Fig. 5: Moves of the semantic game of counterfactuals

Each symbol of the counterfactual logic is resolved through one or more resolution steps. Take (9) and (10) for the resolution of the necessity operator as an example. The operator is resolved in two steps by first making the move (9) and then (10). Each such sequence of resolution steps both begins and ends with a game state of the form $(\varphi, w)_p$. The meaning of this type of game state is that the active player p attempts to show the formula φ at the world w through subsequent play. Accordingly, the game's moves are defined in such a way that the player p wins the game with optimal play, iff φ is true at w. The formula φ is resolved top-down, by continually resolving the weakest-binding operator or atomic formula, until no further resolution is possible. Since the truth conditions of the operators of counterfactual logic depend on the truth of the subformulas they bind, a game to determine the truth of the relevant subformula—or in some cases a different formula—is played after each resolution of a counterfactual operator. For this reason all resolution sequences of all operators and atomic formulas end with the type of game state $(\varphi, w)_p$ that describes the initial state of a new game pertaining to the truth of the formula φ. This allows us to consider the resolution sequences for each operator separately. Given the invariant that the player p wins the game with optimal play from the game state $(\varphi, w)_p$, iff φ is true at w for each subformula φ. Because players frequently switch the role of active player throughout resolution sequences, it will become cumbersome and confusing to describe players by calling them the active or inactive player. We will thus call the player that is active at the beginning of a resolution sequence the proving player and their opponent the disproving player.

(1) Atomic formulas which are true at the current world are resolved to a $\top$-symbol. The proving player subsequently reaches the $\top$-symbol and wins the game. In any game state where the current formula is an atomic formula that is not true at the current world the proving player loses since no move is possible and they are stuck.

(2) The negation is resolved by switching the active player. The reason for this is that the truth of the formula $\neg\varphi$ can be disproven by proving the formula φ. Thus this move switches the active player and makes the disproving player prove the subformula φ. If they are able to do so, they win and the proving player loses. Otherwise they lose and the proving player wins the game.

(3), (4) The disjunction is resolved through the proving player's choice of which subformula to prove i.e. the choice to make the move (3) or (4). If one of the subformulas is true, then the proving player can choose it, prove it through further play and win. In any other case the proving player loses with further optimal play.

(5), (6), (7) Conjunctions are resolved by making the disproving player temporarily become the active player and resolve the conjunction as if it were a disjunction. In other words, the disproving player becomes active and chooses which subformula the proving player has to prove through play, before becoming inactive again. With optimal play the disproving player chooses the harder subformula for their opponent to prove. If one of the subformulas cannot be proven through play, the disproving player chooses it and wins subsequently with optimal play. Conversely if such a subformula does not exist.

(8) The possibility operator is resolved by having any accessible world become the new current world. The proving player gets to choose that world and has to subsequently prove through play that the subformula φ is true at it. If there exists any accessible φ-world, then the proving player can choose that world, show φ there through play and win the game. If no such world exists, the proving player has to choose a world φ cannot be proven at and loses subsequently with optimal play.

(9), (10) The resolution of the necessity operator relates to the resolution of the possibility operator in a similar way as the resolution of conjunctions relates to the resolution of

disjunctions. Both employ a paradigm where the relationship of duality with another operator is used by switching the active player and resolving it like their respective counterpart operator. First the disproving player becomes temporarily active (9), resolves necessity as if it were possibility (10), becomes inactive again and leaves the proving player to prove the subformula φ at the chosen world. This is because necessity is true, when it's supposition is true at all accessible worlds. Thus the disproving player makes the worst choice in the proving player's stead. If there exists any accessible world at which the subformula φ is not true, it is sure to be chosen with optimal play. The proving player then cannot prove φ there through play and loses the game. If φ is true at every accessible world however, then no world can be chosen, such that the proving player cannot prove φ there through play. The proving player wins the game with optimal play in that case.

(11), (12), (13), (14) Sticking closely to the structure of Lewis' definition the counterfactual would's truth conditions, the proving player first decides which truth condition—vacuous or non-vacuous truth—to claim truth of and force their opponent to attempt to disprove. Move type (11) is the choice of vacuous truth and (12) non-vacuous truth. When the proving player claims vacuous truth, they are required to prove the formula $\Box\neg\varphi$. This means that every accessible world has to not be a φ-world. If an accessible φ-world exists, the disproving player can choose it as per $\Box$, become the active player as per $\neg$, then prove φ there and win the game. If no such world exists, the disproving player will fail to prove φ with optimal play and lose.

When deciding to claim that the counterfactual would in question is non-vacuously true, the initial active player also chooses a sphere of accessibility around the current world that is meant to prove non-vacuous truth. This sphere has to contain at least one φ-world and throughout each world within it $\varphi \rightarrow \psi$ has to hold. For the semantic game we adopt a slightly different but unequivocally equivalent condition. We give a sphere around a world i through a world that is a most dissimilar world to i still contained within the sphere. We call such a world a sphere-delimiting or delimiting world. We say that the chosen sphere's delimiting world has to be a φ-world and that $\varphi \rightarrow \psi$ has to hold throughout each world of that sphere. We know that whenever there is a sphere containing a φ-world throughout which $\varphi \rightarrow \psi$ holds, then there also is a subset of that sphere that is a sphere with a delimiting φ-world throughout which $\varphi \rightarrow \psi$ holds. We can simply take the φ-world of the former sphere as the delimiting world for the latter. And whenever there is a sphere with a delimiting φ-world throughout which $\varphi \rightarrow \psi$ holds, then that sphere is also a sphere containing a φ-world throughout which $\varphi \rightarrow \psi$ holds. Now that we have introduced the semantic game's notion of the counterfactual would's non-vacuous truth, let us explain the corresponding moves (13) and (14). After the proving player's choice to claim that the counterfactual is non-vacuously true, the disproving player chooses in which way to attempt to disprove that claim. They can claim that the chosen sphere-delimiting world is not a φ-world (13) and force the initial active player to subsequently prove it through play. Or they can assert that $\varphi \rightarrow \psi$ does not hold throughout the chosen sphere (14) and choose a world serving as a counterexample. The proving player then has to prove that $\varphi \rightarrow \psi$ holds at that world by proving through play that $\neg\varphi \vee \psi$ holds at it.

(15), (16), (17), (18), (19) In a similar vain to the conjunction and necessity, the counterfactual might's resolution hinges on it's interdefinability with the counterfactual would. Since the interdefinition of the counterfactual would and might contains two negations however, it's resolution sequence does not show such a clear correspondence of the counterfactual would's. The counterfactual might is resolved by switching the active player temporarily (15) until the choice between (16) and (17) has been made and follows the same structure as the resolution sequence of the counterfactual would otherwise.

5.5 Limit Assumption

Additionally, Lewis states the limit assumption which allows a simplification of the semantic game of counterfactuals. We will call a sphere that contains a φ-world a φ-permitting sphere.

Definition 8 (Limit assumption)**.** For every world i and formula φ for which a φ-permitting sphere around i exists, there is a smallest φ-permitting sphere around i.

In other words, a nonempty set of φ-permitting spheres around i has a smallest member. Without a doubt this is true when there are only finitely many spheres around i. We can reassure us of that fact the following way. Given any sphere of a nonempty set of φ-permitting spheres $\$_i$ around a world i, we can check for each sphere in $\$_i$ whether it is smaller than that sphere. If that is the case for any sphere, we can take this sphere as our next sphere and repeat the same process for it. If we do not find a smaller sphere, then we found a smallest sphere. And since we assumed $\$_i$ to be finite, we will invariably run out of smaller spheres to find. If we take $\$_i$ to be an infinite set however, we may find an infinite descending sequence of smaller and smaller spheres without end. In that case, we cannot find a smallest sphere, since every sphere has a sphere that is smaller than it.

Since the spheres around i are nested, one of any two spheres contains the other. It follows that the smallest sphere of a set of spheres $\$_i$, if it exists, is the intersection of all spheres of $\$_i$. This means that it is contained in all other spheres in $\$_i$. This circumstance proves quite useful to reduce the semantic game's complexity. This is because showing that a formula φ does not hold throughout a smallest sphere around a world i also shows that it does not hold throughout any larger sphere and by extension any sphere in $\$_i$. We will apply this notion to simplify the semantic game's resolution of counterfactual operators and thereby provide an alternative formulation of it by employing the limit assumption.

5.6 Alternative moves

Definition 9 (Simplified moves)**.** $\rightarrow_l$ is a set of moves, such that it contains exactly those tuples that fit any of the forms described by the following 17 types of moves. Variables are quantified the same way as in definition 7.

$$(x, w)_p \xrightarrow{[x \in F(w)]} (\top, w)_p \tag{1}$$

$$(\neg\varphi, w)_p \rightarrow (\varphi, w)_{np} \tag{2}$$

$$(\varphi \vee \psi, w)_p \rightarrow (\varphi, w)_p \tag{3}$$

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

$$(\varphi \wedge \psi, w)_p \rightarrow (\varphi \wedge \psi, w, p)_{np} \tag{5}$$

$$(\varphi \wedge \psi, w, e)_p \rightarrow (\varphi, w)_e \tag{6}$$

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

$$(\Diamond\varphi, w)_p \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi, w')_p \tag{8}$$

$$(\Box\varphi, w)_p \rightarrow (\Box\varphi, w, p)_{np} \tag{9}$$

$$(\Box\varphi, w, e)_p \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi, w')_e \tag{10}$$

$$(\varphi \diamond\!\!\rightarrow \psi, w)_p \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi \diamond\!\!\rightarrow \psi, w, w', r)_{np} \tag{11}$$

$$(\varphi \diamond\!\!\rightarrow \psi, w, w', r)_p \rightarrow (\varphi \wedge \psi, w')_{np} \tag{12}$$

$$(\varphi \diamond\!\!\rightarrow \psi, w, w', r)_p \xrightarrow{[w \overset{r^*}{\rightsquigarrow} w^*, r^* < r]} (\neg\varphi, w^*)_{np} \tag{13}$$

$$(\varphi \Box\!\!\rightarrow \psi, w)_p \rightarrow (\varphi \Box\!\!\rightarrow \psi, w, p)_{np} \tag{14}$$

$$(\varphi \Box\!\!\rightarrow \psi, w, e)_p \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi \Box\!\!\rightarrow \psi, w, w', r)_e \tag{15}$$

$$(\varphi \Box\!\!\rightarrow \psi, w, w', r)_p \rightarrow (\neg\varphi \vee \psi, w')_p \tag{16}$$

$$(\varphi \Box\!\!\rightarrow \psi, w, w', r)_p \xrightarrow{[w \overset{r^*}{\rightsquigarrow} w^*, r^* < r]} (\varphi, w^*)_p \tag{17}$$

Fig. 6: Moves employing the limit assumption

Moves (1)-(10) are the same as in the formulation in figure 5.

(11), (12), (13) With respect to the limit assumption, we can formulate a simpler resolution of the counterfactual might. It can be proven by showing that the smallest φ-permitting sphere around the current world w does contain a $(\varphi \wedge \psi)$-world. In that case every larger sphere around w also contains a $(\varphi \wedge \psi)$-world; and every smaller sphere around w does not contain a φ-world. This corresponds to the requirement (2) of Lewis' truth conditions of the counterfactual might (see figure 4). Furthermore, it is necessary to show that a φ-permitting sphere around w exists. Hence the counterfactual might's resolution sequence begins with the proving player's choice of a sphere of accessibility (11). When the proving player cannot choose a sphere, because no other world is accessible, they lose. In that case, no φ-permitting sphere around w exists. In the case that the proving player is able to choose a sphere-delimiting world and thus a sphere, the chosen world ought to be a $(\varphi \wedge \psi)$-world. This is because we combine two truth requirements that the sphere needs to contain a φ-world; and that the sphere needs to contain a $(\varphi \wedge \psi)$-world. If the latter is true, the former is

true as well. Thus we can omit the less stringent requirement. Analogously to the previous formulation of the counterfactual would's resolution, we assert without loss of generality that the chosen sphere be $(\varphi \wedge \psi)$-delimited instead of merely containing a $(\varphi \wedge \psi)$-world. If the proving player did not choose a $(\varphi \wedge \psi)$-world, then the disproving player is able to refute that by forcing the proving player to prove $(\varphi \wedge \psi)$ at the chosen world (12). If the delimiting world is either not a φ-world or not $(\varphi \wedge \psi)$-world, then the proving player loses subsequently with optimal play. And if there exists a smaller φ-permitting sphere than the sphere chosen by the proving player, then that players opponent can choose a closer world and force the proving player to attempt to show through play that the chosen world is not a φ-world (13). If it is a φ-world, then the proving player loses with optimal play and wins otherwise.

(14), (15), (16), (17) The counterfactual would is resolved by switching the active player before and after the choice of a sphere of accessibility and resolving it similarly to a counterfactual might. First, the active player is switched (14). Then the disproving player chooses a sphere of accessibility (15). And finally the proving player becomes active again and chooses in which way to disprove the disproving players chosen sphere. The proving player may either prove that $\varphi \rightarrow \psi$ holds at the chosen sphere-delimiting world by proving that $\neg\varphi \vee \psi$ holds there (16) or they may prove that there is a smaller φ-permitting sphere by proving that there is a more similar φ-world (17).

6 Playing the game

In order to describe the course of a played game we define plays.

Definition 10 (Plays)**.** Given the semantic game $(I, E, (G, R), S)$, a sequence of game states $P = (p_0, p_1, p_2, ...)$ is called a (winning) play if and only if the following conditions hold.

(1) P begins at the semantic game's initial game state; that is, $p_0 = I$.

(2) P is consistent; that is, for each pair of game states (p_i, p_{i+1}) with $i \in \mathbb{N}$ of which both members are elements of P it holds that $(p_i, p_{i+1}) \in R$.

(3) P is finished; that is, P has a last element p_n and either $p_n \in E$ or $(\{p_n\} \times G) \cap R = \emptyset$.

(W) P is winning for a player p; that is, (3) is true and it either holds that $p_n \in E$ and p is the active player at p_n or that $(\{p_n\} \times G) \cap R = \emptyset$ and p is not the active player at p_n.

Plays are sequences of game states and describe the order in which game positions are reached in an instance of play of the semantic game of counterfactuals. A play can thus be thought of as a history of game states that is dependent on the decisions of players and the concrete semantic game that is played. Since there are are many differing instances of the semantic game, varying in the initial formula, initial world, truth of atomic formulas at worlds and the similarity relation between worlds, it is clear that a play is only a play relative to a certain semantic game, since it has to fit the constraints imposed by that game's rules. It may well be the case that a play is a play relative to one instance of the semantic game, but not another. To be called a play with respect to a semantic game of counterfactuals, a sequence of game states has to (1) begin at the initial game state of that game, (2) only make transitions between game states that are moves of the game and (3) end with a game state where one player wins. If (3) is not true, but (1) and (2) are, we call P an unfinished or partial play. Additionally, we call a play winning for a player, if that player reaches the $\top$-symbol at the end of the play or is not the stuck player at the end of the play. (W)

Definition 11 (Strategies)**.** Given the transition system (G, R), a strategy $S : G \to R$ is a partial mapping from game states to moves.

- two strategies produce a play - strategy winning, when the produces play is winning for the player employing it - strategy (mapping from game states to moves (positional game, thus state-based is okay, otherwise partial mapping from partial plays to moves))

6.1 Correctness proof

Show inductively for every formula that the corresponding game is correct

6.2 Termination proof

Show that every resolution sequence monotonously reduces the current formulas size - the game terminates -> therefore it is a decided game (Every game is won by one player)

7 The educational game

7.1 Correspondance between semantic game and computer game

7.1.1 Limit assumption justification

- better gameplay & usability
- no multiple click resolution
- less complex operators
- finite graph satisfies it anyways, nicer interactive presentation - no weird negation magic

7.2 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
```

8 Results

8.1 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.