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 that it always halts after a finite number of moves; (iii) prove its correctness; (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 $\square\!\!\rightarrow$ and counterfactual might $\diamond\!\!\rightarrow$ operators as binary modal operators. [Lew73] When we write $\varphi \square\!\!\rightarrow \psi$, 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 $\square\!\!\rightarrow$ I was punctual" and read it the following way. Read $\varphi \square\!\!\rightarrow \psi$ 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.
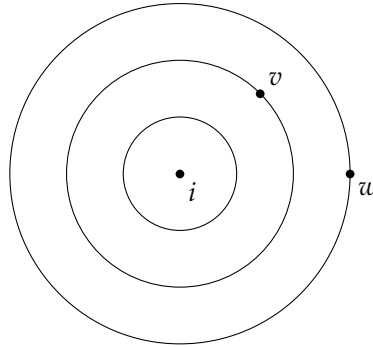

Fig. 2: A set of spheres $\$_i = \{\{i\}, \{i, v\}, \{i, v, w\}\}$

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.

$\varphi \; \Box\!\!\rightarrow \psi$ 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

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 $\square(\varphi \rightarrow \psi)$ 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.

$$\text{"If Kennedy had pressed the red button, there would have been nuclear war"} \quad (1)$$

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.

$\varphi \diamondsuit\rightarrow \psi$ 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.

$$\varphi \square\rightarrow \psi = \neg(\varphi \diamondsuit\rightarrow \neg\psi) \quad (2)$$

$$\varphi \diamondsuit\rightarrow \psi = \neg(\varphi \square\rightarrow \neg\psi) \quad (3)$$

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

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

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 $(W, \rightsquigarrow, F)$, consisting of

- the set of all possible worlds $W = \{w, v, ...\}$,

- the similarity relation $\rightsquigarrow: W \times \mathbb{R} \times W$ with $\{(w, 0, w) \mid w \in W\} \subseteq \rightsquigarrow$, and

- an assignment of each world, to a set of atomic formulas $F: W \to 2^{Atoms}$.

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 $S, w' \vDash \varphi$.

$S, w \vDash \varphi \;\Box\!\!\rightarrow\; \psi$, 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 $S, w^* \vDash \neg\varphi \vee \psi$.

$S, w \vDash \varphi \;\Diamond\!\!\rightarrow\; \psi$, 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 change the active player and current world.

5.1 Definition of the semantic game

The game itself 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 $\mathfrak{S}[\varphi, w, S] = (I, E, G, R)$ with respect to a formula φ, a world w and a counterfactual Kripke structure $S = (W, \rightsquigarrow, F)$ is an ordered tuple, consisting of

- an initial game state $I = (\varphi, w)_d$,

- a set of winning game states $E = \{(\top, w)_p \mid w \in W \wedge p \in \{a, d\}\}$,

- a set of all game states G, and

- a transition relation $R : G \times G$.

The notation $\mathfrak{S}_p[\varphi, w, S]$ means that the annotated player $p \in \{a, d\}$ begins the game; that is, $I = (\varphi, w)_p$ instead of $I = (\varphi, w)_d$.

The initial game state I varies from game to game. It contains the formula φ that ought to be proven and the world w 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 contains all game states in which the active player wins. When the active player wins, the respective other player loses. The set E 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 is stuck. When this happens, the active player loses and the active player's opponent i.e. the inactive player wins. $\mathfrak{S}[\varphi, w, S]$ also 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 $\rightarrowtail$ and one simplified formulation $\twoheadrightarrow$ employing the limit assumption. One may take either as R. In order to arrive at our definition for $\rightarrowtail$ and $\twoheadrightarrow$, let us first define game states and transitions.

5.2 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.

Definition 4 (Game states). A game state $g \in G$ 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 $x, y, \ldots$ and the propositional symbols $\top, \neg, \vee, \diamondsuit$; 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.

5.3 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 5 (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 $\rightarrowtail$ specific to the semantic game of counterfactuals.

Definition 6 (Moves). $\rightarrowtail$ 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; $x \in \textit{Atoms}$ 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 $r, r^* \in \mathbb{R}$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $\square \neg \varphi$. 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 $\square$, 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 $\varphi \to \psi$ 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 $\varphi \to \psi$ has to hold throughout each world of that sphere. We know that whenever there is a sphere containing a φ-world throughout which $\varphi \to \psi$ holds, then there also is a subset of that sphere that is a sphere with a delimiting φ-world throughout which $\varphi \to \psi$ 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 $\varphi \to \psi$ holds, then that sphere is also a sphere containing a φ-world throughout which $\varphi \to \psi$ 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 $\varphi \to \psi$ does not hold throughout the chosen sphere (14) and choose a world serving as a counterexample. The proving player then has to prove that $\varphi \to \psi$ holds at that world by proving through play that $\neg \varphi \lor \psi$ holds at it.

(15), (16), (17), (18), (19) In a similar vein 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.

5.4 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 7 (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.

5.5 Moves under the limit assumption

Definition 8 (Simplified moves). $\rightarrowtail$ 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 6. And the moves (1)-(10) are adopted from definition 6.

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

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

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

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

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

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

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

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

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

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

$$(\varphi \Diamond\!\!\to \psi, w)_p \xhookrightarrow{[w \overset{r}{\leadsto} w']} (\varphi \Diamond\!\!\to \psi, w, w', r)_{np} \tag{11}$$

$$(\varphi \Diamond\!\!\to \psi, w, w', r)_p \rightarrowtail (\varphi \wedge \psi, w')_{np} \tag{12}$$

$$(\varphi \Diamond\!\!\to \psi, w, w', r)_p \xhookrightarrow{[w \overset{r^*}{\leadsto} w^*, r^* < r]} (\neg \varphi, w^*)_{np} \tag{13}$$

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

$$(\varphi \Box\!\!\to \psi, w, e)_p \xhookrightarrow{[w \overset{r}{\leadsto} w']} (\varphi \Box\!\!\to \psi, w, w', r)_e \tag{15}$$

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

$$(\varphi \Box\!\!\to \psi, w, w', r)_p \xhookrightarrow{[w \overset{r^*}{\leadsto} w^*, r^* < r]} (\varphi, w^*)_p \tag{17}$$

Fig. 6: Moves employing the limit assumption

(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 9 (Plays). Given a semantic game of counterfactuals (I, E, G, R), a path of game states $P = p_0 p_1 \ldots \in G^\infty$ 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 every subpath $p_i p_{i+1}$ 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 paths 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. We know that there are many differing instances of the semantic game, varying in the initial formula, initial world and counterfactual Kripke structure the game is played on. Thus a play is a play with respect to a certain semantic game. It has to fit the constraints imposed by a game's rules to be called a play of that game. It may well be the case that a play is a play with regard to one instance of the semantic game, but not another. To be called a play with respect to a semantic game of counterfactuals, a path 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) We also define the notion of strategies.

Definition 10 (Strategies). Given a semantic game of counterfactuals (I, E, G, R) a positional strategy $T \subseteq \{(g_1, (g_1, g_2)) \mid g_1, g_2 \in G \wedge (g_1, g_2) \in R\}$ is a partial mapping from game states to moves. A strategy T_p of a player p is called consistent with a play P, iff for every subpath $p_i p_{i+1}$ of P where the player p is the active player at p_i, T_p maps p_i to (p_i, p_{i+1}). We call a strategy T_p of a player p winning, iff each play that T_p is consistent with is a winning play for player p. A player that has a winning strategy for a semantic game is said to win that semantic game.

One can intuit the meaning of a strategy to be a declaration of intent of which moves a player will plays in which game position. We codify a strategy as a partial mapping from game states to moves and call it a positional strategy, because it is state-based and does not take into account any previous game history i.e. a partial play. Since the semantic game's win conditions are independent from previous play, it suffices to define strategies this way. Additionally, the previously mentioned partial mapping also underlies the constraint that a game state g has to be mapped to a move with g as its initial game state. This way a strategy can only map a game state to moves that are possible at it. Furthermore we introduced the notion of consistency with a play. Given a player p and their strategy T_p, that strategy is called consistent with a play P, iff at each game state g of P where p made a move, that move is the same as $T_p(g)$. Colloquially we might say "The player p followed the strategy T_p in the play P". Thus we can say that following a strategy T_p means that the set of plays that can occur is restricted to the set of plays T_p is consistent with. If and only if each of those plays is a winning play for player p, then the strategy T_p is winning, because the opponent of the player p is unable to make moves such that p loses. Note that it is clearly impossible for both players to have a winning strategy with respect to the same game at the same time. If one player has a winning strategy, their opponent does not.

6.1 Termination proof

Let us show that the semantic game of counterfactuals $\mathfrak{S}[\varphi, w, S] = ((\varphi, w)_p, E, G, \rightarrowtail)$ ends. Whenever a player cannot make a move and is stuck, that player loses and the game ends. We show that at some point either a player becomes stuck or a player reaches a winning game state. We begin by showing that all games with a current counterfactual formula with formula height 1 end. And we follow up by showing that every resolution sequence for every operator reduces the formula height or ends the game. We define the height function for counterfactual formulas.

Definition 11 (Height function).

$$h(\varphi) = \begin{cases} 1, \text{ for } \varphi \in Atoms \cup \{\top, \bot\} \\ h(\varphi') + 1, \text{ for } \varphi = \bigcirc\varphi' \text{ and } \bigcirc \in \{\neg, \Box, \Diamond\} \\ max(h(\varphi'), h(\psi')) + 1, \text{ for } \varphi = \varphi' \bigcirc \psi' \text{ and } \bigcirc \in \{\vee, \wedge, \Box\rightarrow, \Diamond\rightarrow\} \end{cases}$$

$$max(x, y) = \begin{cases} x, \text{ for } x \geq y \\ y, \text{ for } x < y \end{cases}$$

Proof. We show that the games with formulas with formula height 1 end.

- Base Case $((\top, w)_p, E, G, \rightarrowtail)$:
  $(\top, w)_p \in E$ is a winning position for the player p. The game ends without further play.

- Base Case $((\bot, w)_p, E, G, \rightarrowtail)$:

  $\rightarrowtail$ does not contain any move with $(\bot, w)_p$ as its first game state. The player p cannot play a move, is stuck and loses. The game ends without further play.

- Base Case $((x, w)_p, E, G, \rightarrowtail)$ with $x \in Atoms$:

$$(x, w)_p$$
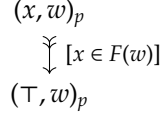$$\downarrow [x \in F(w)]$$
$$(\top, w)_p$$

Fig. 7: resolution of atomic formulas

If $x \notin F(w)$, then $\rightarrowtail$ does not contain any move with $(x, w)_p$ as its first game state. The player p cannot play a move, is stuck and loses. The game ends.

If $x \in F(w)$, then $\rightarrowtail$ contains a single move $((x, w)_p, (\top, w)_p)$ with $(x, w)_p$ as its first game state. The player p cannot play any other move and wins by reaching the winning position $(\top, w)_p \in E$. The game ends.

We show that the formula height is reduced by all other resolution sequences.
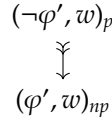
- Case $((\neg \varphi', w)_p, E, G, \rightarrowtail)$:

$$(\neg \varphi', w)_p$$
$$\downarrow$$
$$(\varphi', w)_{np}$$

Fig. 8: resolution of negations

$\rightarrowtail$ contains a single move $((\neg \varphi', w)_p, (\varphi', w)_{np})$ with $(\neg \varphi', w)_p$ as its first game state and $np \in \{a, d\} \setminus \{p\}$. The player p cannot play any other move and reduces the formula height by 1, because $h(\neg \varphi') = h(\varphi') + 1$.

- Case $((\varphi' \vee \psi', w)_p, E, G, \rightarrowtail)$:
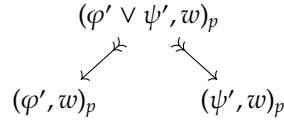
$$(\varphi' \vee \psi', w)_p$$

$$(\varphi', w)_p \qquad (\psi', w)_p$$

Fig. 9: resolution of disjunctions

$\rightarrowtail$ contains two moves $((\varphi' \vee \psi', w)_p, (\varphi', w)_p)$ and $((\varphi' \vee \psi', w)_p, (\psi', w)_p)$ with the game state $(\varphi' \vee \psi', w)_p$ as their first member. The player p can choose which move to play and reduces the formula height by 1 or more, because $h(\varphi' \vee \psi') = max(h(\varphi'), h(\psi')) + 1$.

17

- Case $((\varphi' \wedge \psi', w)_p, E, G, \rightarrowtail)$:

$$(\varphi' \wedge \psi', w)_p$$
$$\downarrow$$
$$(\varphi' \wedge \psi', w, p)_{np}$$
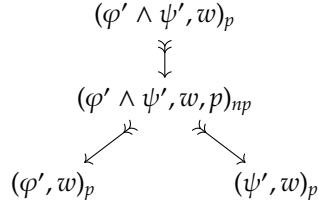$$(\varphi', w)_p \qquad\qquad (\psi', w)_p$$

Fig. 10: resolution of conjunctions

$\rightarrowtail$ contains a single move $((\varphi' \wedge \psi', w)_p, (\varphi' \wedge \psi', w, p)_{np})$ with $(\varphi' \wedge \psi', w)_p$ as its first game state and $np \in \{a, d\} \setminus \{p\}$. The player p cannot play any other move and does so. Then $\rightarrowtail$ contains two moves $((\varphi' \wedge \psi', w, p)_{np}, (\varphi', w)_p)$ and $((\varphi' \wedge \psi', w)_{np}, (\psi', w)_p)$ with the game state $(\varphi' \wedge \psi', w, p)_{np}$ as their first member. The player np chooses a move to play and reduces the formula height by 1 or more, because $h(\varphi' \wedge \psi') = max(h(\varphi'), h(\psi')) + 1$.
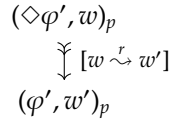
- Case $((\Diamond\varphi', w)_p, E, G, \rightarrowtail)$:

$$(\Diamond\varphi', w)_p$$
$$\downarrow \; [w \overset{r}{\rightsquigarrow} w']$$
$$(\varphi', w')_p$$

Fig. 11: resolution of possibility

$\rightarrowtail$ contains a move of the form $((\Diamond\varphi', w)_p, (\varphi', w')_p)$ with $(\Diamond\varphi', w)_p$ as its first game state for each world $w' \in W_w$. The player p can choose to play one of those moves and reduces the formula height by 1, because $h(\Diamond\varphi') = h(\varphi') + 1$.

- Case $((\Box\varphi', w)_p, E, G, \rightarrowtail)$:

$$(\Box\varphi', w)_p$$
$$\downarrow$$
$$(\Box\varphi', w, p)_{np}$$
$$\downarrow \; [w \overset{r}{\rightsquigarrow} w']$$
$$(\varphi', w')_p$$
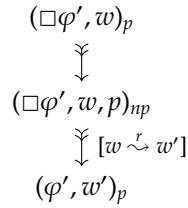
Fig. 12: resolution of necessity

$\rightarrowtail$ contains a single move $((\Box\varphi', w)_p, (\Box\varphi', w, p)_{np})$ with $(\Box\varphi', w)_p$ as its first game state and $np \in \{a, d\} \setminus \{p\}$. The player p cannot play any other move and does so. Then

18

$\twoheadrightarrow$ contains a move of the form $((\Box\varphi', w, p)_{np}, (\varphi', w')_p)$ with $(\Box\varphi', w, p)_{np}$ as its first game state for each world $w' \in W_w$. The player np can choose to play one of those moves and reduces the formula height by 1, because $h(\Box\varphi') = h(\varphi') + 1$.

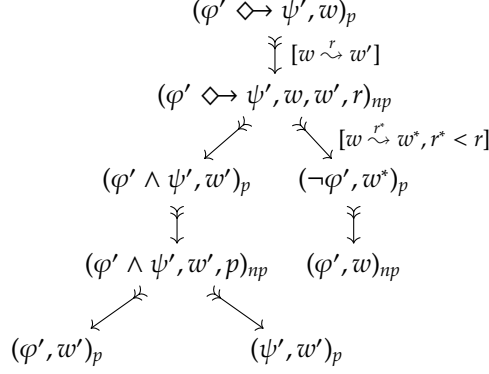- Case $((\varphi' \Diamond\!\!\rightarrow \psi', w)_p, E, G, \twoheadrightarrow)$:

$$(\varphi' \Diamond\!\!\rightarrow \psi', w)_p$$

$$\big\uparrow \ [w \overset{r}{\leadsto} w']$$

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

$$\swarrow \qquad \searrow \ [w \overset{r}{\leadsto} w^*, r^* < r]$$

$$(\varphi' \wedge \psi', w')_p \qquad (\neg\varphi', w^*)_p$$

$$(\varphi' \wedge \psi', w', p)_{np} \qquad (\varphi', w)_{np}$$

$$(\varphi', w')_p \qquad\qquad (\psi', w')_p$$

Fig. 13: resolution of the counterfactual might

$\twoheadrightarrow$ contains a move of the form $((\varphi' \Diamond\!\!\rightarrow \psi', w)_p, (\varphi' \Diamond\!\!\rightarrow \psi', w, w', r)_{np})$ with the game state $(\varphi' \Diamond\!\!\rightarrow \psi', w)_p$ as its first member and $np \in \{a, d\} \setminus \{p\}$ for each world $w' \in W_w$, $r \in \mathbb{R}$ and $w \overset{r}{\leadsto} w'$. The player p chooses to play one of those moves. Then $\twoheadrightarrow$ contains a move of the form $((\varphi' \Diamond\!\!\rightarrow \psi', w, w', r)_{np}, (\neg\varphi', w^*)_p)$ with the game state $((\varphi' \Diamond\!\!\rightarrow \psi', w, w', r)_{np}$ as its first member for each $w^* \in W_w$; $r^* \in \mathbb{R}$ with $w \overset{r}{\leadsto} w^*$ and $r^* < r$; and contains the move $((\varphi' \Diamond\!\!\rightarrow \psi', w, w', r)_{np}, (\varphi' \wedge \psi', w')_p)$. If the player np chooses and plays one of the former moves, then the formula height is reduced by 1, because $h(\varphi' \Diamond\!\!\rightarrow \psi') = max(h(\varphi'), h(\psi')) + 1$. If the player np chooses and plays the latter and the player p either plays resolves the resulting conjunction afterwards, then the formula height is reduced by 1, because $h(\varphi' \Diamond\!\!\rightarrow \psi') = max(h(\varphi'), h(\psi')) + 1$.

- Case $((\varphi' \Box\!\!\rightarrow \psi', w)_p, E, G, \twoheadrightarrow)$:
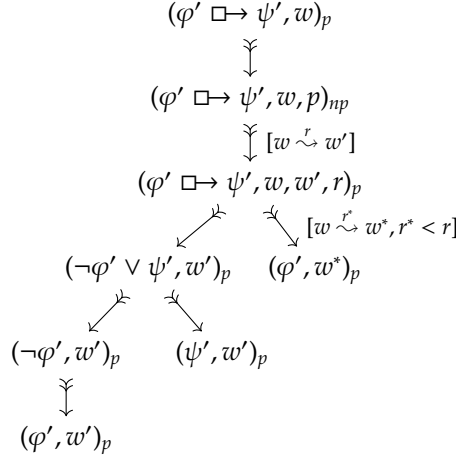
$$(\varphi' \mathrel{\Box\!\!\rightarrow} \psi', w)_p$$

$$(\varphi' \mathrel{\Box\!\!\rightarrow} \psi', w, p)_{np}$$

$$[w \overset{r}{\leadsto} w']$$

$$(\varphi' \mathrel{\Box\!\!\rightarrow} \psi', w, w', r)_p$$

$$[w \overset{r^*}{\leadsto} w^*, r^* < r]$$

$$(\neg\varphi' \lor \psi', w')_p \qquad (\varphi', w^*)_p$$

$$(\neg\varphi', w')_p \qquad (\psi', w')_p$$

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

Fig. 14: resolution of the counterfactual would

$\twoheadrightarrow$ contains a single move $((\varphi' \mathrel{\Box\!\!\rightarrow} \psi', w)_p, (\varphi' \mathrel{\Box\!\!\rightarrow} \psi', w, p)_{np})$ with $(\varphi' \mathrel{\Box\!\!\rightarrow} \psi', w)_p$ as its first game state and $np \in \{a, d\} \setminus \{p\}$. The player p cannot play any other move and does so. $\twoheadrightarrow$ contains a move of the form $((\varphi' \mathrel{\Box\!\!\rightarrow} \psi', w, p)_{np}, (\varphi' \mathrel{\Box\!\!\rightarrow} \psi', w, w', r)_p)$ with the game state $(\varphi' \mathrel{\Box\!\!\rightarrow} \psi', w, p)_{np}$ as its first member for each world $w' \in W_w, r \in \mathbb{R}$ and $w \overset{r}{\leadsto} w'$. The player np chooses to play one of those moves. Then $\twoheadrightarrow$ contains a move of the form $((\varphi' \mathrel{\Box\!\!\rightarrow} \psi', w, w', r)_p, (\varphi', w^*)_p)$ with the game state $((\varphi' \mathrel{\Box\!\!\rightarrow} \psi', w, w', r)_p$ as its first member for each $w^* \in W_w$; $r^* \in \mathbb{R}$ with $w \overset{r^*}{\leadsto} w^*$ and $r^* < r$; and contains the move $((\varphi' \mathrel{\Box\!\!\rightarrow} \psi', w, w', r)_p, (\neg\varphi' \lor \psi', w')_p)$. If the player p chooses and plays one of the former moves, then the formula height is reduced by 1, because $h(\varphi' \mathrel{\Box\!\!\rightarrow} \psi') = max(h(\varphi'), h(\psi')) + 1$. If the player np chooses and plays the latter and subsequently resolves the resulting disjunction, then the formula height of the resulting current formula is reduced by 1, because $h(\varphi' \mathrel{\Box\!\!\rightarrow} \psi') = max(h(\varphi'), h(\psi')) + 1$.

Considering that all games with a counterfactual formula of formula height 1 end, every counterfactual formula has a formula height greater than 0 and every resolution sequence of a counterfactual operator reduces the height of the current formula of a game by 1, all semantic games end. □

Since a semantic game can only result in a victory for for either the defender or attacker, the semantic game of counterfactuals is decided. If every game ends and the game results in the victory of one player, then every game is won by one player.

6.2 Correctness proof

We will show through induction that the semantic game of counterfactuals is correct. More precisely, we will show that for each formula φ, world w and counterfactual Kripke structure $S = (W, \leadsto, F)$, the defending player has a winning strategy in the corresponding semantic game $\mathfrak{S}[\varphi, w, S]$, if and only if $S, w \vDash \varphi$. Because the defender always begins a game $\mathfrak{S}[\varphi, w, S]$ as the active player we will show that the player p wins the game $\mathfrak{S}_p[\varphi, w, S]$, iff $S, w \vDash \varphi$. For this proof, we take the set of moves $\twoheadrightarrow$ as the set of game state transitions R.

Proof. We begin by showing 3 base cases.

- Base Case $\varphi = \top$:
  $S, w \vDash \top$ is true and the respective semantic game is $\mathfrak{S}_p[\top, w, S] = ((\top, w)_p, E, G, \rightarrowtail l)$. Since $(\top, w)_p \in E$, it follows that the player p wins the game with any strategy.

- Base Case $\varphi = \bot$:
  $S, w \vDash \bot$ is false and the semantic game is $\mathfrak{S}_p[\bot, w, S] = ((\bot, w)_p, E, G, \rightarrowtail)$. There is no move in $\rightarrowtail$ with $(\bot, w)_p$ as its first element (see fig.6). That means no move can be played in this game state. The player p is stuck and loses the game with any strategy.

- Base Case $\varphi = x$ and $x \in Atoms$:
  $S, w \vDash x$ is true iff $x \in F(w)$. And the semantic game is $\mathfrak{S}_p[x, w, S] = ((x, w)_p, E, G, \rightarrowtail)$. We show that the player p has a winning strategy, iff $x \in F(w)$. Suppose $x \in F(w)$, then $\rightarrowtail$ contains the move $((x, w)_p, (\top, w)_p)$ per definition of $\rightarrowtail$ (see figure 6 move 1). $((x, w)_p, (\top, w)_p)$ is a winning play for the player p. And any strategy mapping $(x, w)_p$ to $((x, w)_p, (\top, w)_p)$ is a winning strategy for the player p. Suppose $x \notin F(w)$, then $\rightarrowtail$ does not contain any move with $(x, w)_p$ as its first element (see figure 6). That means no move can be played. The player p is stuck and loses the game with any strategy.

Given the induction assumption that the player p' wins the game $\mathfrak{S}_{p'}[\varphi', w, S]$, iff $S, w \vDash \varphi'$; and the player p' wins the game $\mathfrak{S}_{p'}[\psi', w, S]$, iff $S, w \vDash \psi'$, we show for each composite formula that the player p wins the game $\mathfrak{S}_p[\varphi, w, S]$, iff $S, w \vDash \varphi$.

- Induction Step $\varphi = \neg\varphi'$:
  $S, w \vDash \neg\varphi'$ is true iff $S, w \nvDash \varphi'$. And the game is $\mathfrak{S}_p[\neg\varphi', w, S] = ((\neg\varphi', w)_p, E, G, \rightarrowtail)$. The only move that has $(\neg\varphi', w)_p$ as its first element and is contained within $\rightarrowtail$ is $((\neg\varphi', w)_p, (\varphi', w)_{np})$ with $np \in \{a, d\} \setminus \{p\}$ (see figure 6). For this reason, every play of the game $\mathfrak{S}_p[\neg\varphi', w, S]$ begins with the partial play $(\neg\varphi', w)_p(\varphi', w)_{np}$ with $np \in \{a, d\} \setminus \{p\}$. And given the induction assumption we know that the player np wins the game $\mathfrak{S}_{np}[\varphi', w, S] = ((\varphi', w)_{np}, E, G, \rightarrowtail)$, iff $S, w \vDash \varphi'$. Since the player p loses when the player np wins and the player p wins when the player np loses, we know that the player p wins, iff $S, w \nvDash \varphi'$.

- Induction Step $\varphi = \varphi' \vee \psi'$:
  $S, w \vDash \varphi' \vee \psi'$ is true, iff ($S, w \vDash \varphi'$ or $S, w \vDash \psi'$). And the corresponding game is $\mathfrak{S}_p[\varphi' \vee \psi', w, S] = ((\varphi' \vee \psi', w)_p, E, G, \rightarrowtail)$. $\rightarrowtail$ contains exactly the two moves $((\varphi' \vee \psi', w)_p, (\varphi', w)_p)$ and $((\varphi' \vee \psi', w)_p, (\psi', w)_p)$ that have $(\varphi' \vee \psi', w)_p$ as their first element. Given the induction assumption, the player p has a winning strategy for the game $\mathfrak{S}_p[\varphi', w, S]$, iff $S, w \vDash \varphi'$; and further has a winning strategy for the game $\mathfrak{S}_p[\psi', w, S]$, iff $S, w \vDash \psi'$. Suppose $S, w \vDash \varphi'$, then the player $p's$ winning strategy for $\mathfrak{S}_p[\varphi', w, S]$ is also a winning strategy for $\mathfrak{S}_p[\varphi' \vee \psi', w, S]$, if it maps $(\varphi' \vee \psi', w)_p$ to $((\varphi' \vee \psi', w)_p, (\varphi', w)_p)$. And suppose $S, w \vDash \psi'$, then the player $p's$ winning strategy for $\mathfrak{S}_p[\psi', w, S]$ is also a winning strategy for $\mathfrak{S}_p[\varphi' \vee \psi', w, S]$, if it maps $(\varphi' \vee \psi', w)_p$ to $((\varphi' \vee \psi', w)_p, (\psi', w)_p)$. Lastly suppose $S, w \nvDash \varphi'$ and $S, w \nvDash \psi'$, then no winning strategy exists for the player p and any of the games $\mathfrak{S}_p[\varphi', w, S]$ and $\mathfrak{S}_p[\psi', w, S]$. Because the player p cannot make any move that does not transition into one of those two games, no winning strategy exists for the player p and game $\mathfrak{S}_p[\varphi' \vee \psi', w, S]$.

- Induction Step $\varphi = \varphi' \wedge \psi'$:
  $S, w \vDash \varphi' \wedge \psi'$ is true, iff ($S, w \vDash \varphi'$ and $S, w \vDash \psi'$). And the corresponding game is $\mathfrak{S}_p[\varphi' \wedge \psi', w, S] = ((\varphi' \wedge \psi', w)_p, E, G, \rightarrowtail)$. Every play of $\mathfrak{S}_p[\varphi' \wedge \psi', w, S]$ begins

with either one of the following partial plays $(\varphi' \wedge \psi', w)_p(\varphi' \wedge \psi', w, p)_{np}(\varphi', w)_p$ and $(\varphi' \wedge \psi', w)_p(\varphi' \wedge \psi', w, p)_{np}(\varphi', w)_p$ with $np \in \{a, d\} \setminus \{p\}$ for each of those partial plays. Given the induction assumption, the player p has a winning strategy for the game $\mathfrak{S}_p[\varphi', w, S]$, iff $S, w \vDash \varphi'$; and further has a winning strategy for the game $\mathfrak{S}_p[\psi', w, S]$, iff $S, w \vDash \psi'$. Suppose $S, w \nvDash \varphi'$, then the player p loses the game $\mathfrak{S}_p[\varphi', w, S]$. In that case, any strategy that maps $(\varphi' \wedge \psi', w, p)_{np}$ to $((\varphi' \wedge \psi', w, p)_{np}, (\varphi', w)_p)$ is a winning strategy for the player np and game $\mathfrak{S}_p[\varphi' \wedge \psi', w, S]$. Analogously suppose $S, w \nvDash \psi'$, then the player p loses the game $\mathfrak{S}_p[\psi', w, S]$. In that case, any strategy that maps $(\varphi' \wedge \psi', w, p)_{np}$ to $((\varphi' \wedge \psi', w, p)_{np}, (\psi', w)_p)$ is a winning strategy for the player np and game $\mathfrak{S}_p[\varphi' \wedge \psi', w, S]$. Lastly suppose $S, w \vDash \varphi'$ and $S, w \vDash \psi'$, then a winning strategy exists for the player p and each of the games $\mathfrak{S}_p[\varphi', w, S]$ and $\mathfrak{S}_p[\psi', w, S]$. Because the player np cannot make any move that does not transition into one of those two games, no winning strategy exists for the player np and game $\mathfrak{S}_p[\varphi' \wedge \psi', w, S]$.

- Induction Step $\varphi = \Diamond \varphi'$:
  $S, w \vDash \Diamond \varphi'$ is true, iff a world $w' \in W_w$ exists, such that $S, w \vDash \varphi'$. And the semantic game is $\mathfrak{S}_p[\Diamond \varphi', w, S] = ((\Diamond \varphi', w)_p, E, G, \rightarrowtail)$. Every play of $\mathfrak{S}_p[\Diamond \varphi', w, S]$ begins with a partial play of the form $(\Diamond \varphi', w)_p(\varphi', w')_p$ where $w' \in W_w$ ranges over the set of worlds W_w that are accessible from w (see figure 6 move 8). Suppose there is an accessible world $w' \in W_w$, such that $S, w' \vDash \varphi'$. Then given the induction assumption, it is the case that the player p has a winning strategy for the corresponding game $\mathfrak{S}_p[\varphi', w', S]$. And if that strategy maps $(\Diamond \varphi', w)_p$ to $((\Diamond \varphi', w)_p, (\varphi', w')_p)$, it is also a winning strategy for the player p and game $\mathfrak{S}_p[\Diamond \varphi', w, S]$. Conversely suppose there is no accessible world $w' \in W_w$, such that $S, w' \vDash \varphi'$. Then given the induction assumption, the player p does not have a winning strategy for any game of the form $\mathfrak{S}_p[\varphi', w', S]$ where $w' \in W_w$. But because every play of the game $\mathfrak{S}_p[\Diamond \varphi', w, S]$ begins with a partial play that transitions into a game of the form $\mathfrak{S}_p[\varphi', w', S]$ where $w' \in W_w$, the player p loses the game $\mathfrak{S}_p[\Diamond \varphi', w, S]$.

- Induction Step $\varphi = \Box \varphi'$:
  $S, w \vDash \Box \varphi'$ is true, iff for every world $w' \in W_w$, it is true that $S, w \vDash \varphi'$. And the semantic game is $\mathfrak{S}_p[\Box \varphi', w, S] = ((\Box \varphi', w)_p, E, G, \rightarrowtail)$. Every play of $\mathfrak{S}_p[\Box \varphi', w, S]$ begins with a partial play of the form $(\Box \varphi', w)_p(\Box \varphi', w, p)_{np}(\varphi', w')_p$ where $w' \in W_w$ ranges over the set of worlds W_w that are accessible from w and $np \in \{a, d\} \setminus \{p\}$ (see fig. 6 moves 9,10). Suppose that there is an accessible world $w' \in W_w$, such that $S, w' \nvDash \varphi'$. Then given the induction assumption, the player np has a winning strategy for the game $\mathfrak{S}_p[\varphi', w', S]$. And if that strategy maps $(\Box \varphi', w, p)_{np}$ to $((\Box \varphi', w, p)_{np}, (\varphi', w')_p)$, it is also a winning strategy for the player np and game $\mathfrak{S}_p[\Box \varphi', w, S]$. Conversely suppose that there is no accessible world $w' \in W_w$, such that $S, w' \nvDash \varphi'$. Then given the induction assumption, the player np does not have a winning strategy for any game of the form $\mathfrak{S}_p[\varphi', w', S]$ where $w' \in W_w$. But because every play of the game $\mathfrak{S}_p[\Box \varphi', w, S]$ begins with a partial play that transitions into a game of the form $\mathfrak{S}_p[\varphi', w', S]$ where $w' \in W_w$, the player np loses the game $\mathfrak{S}_p[\Box \varphi', w, S]$ and p wins.

- Induction Step $\varphi = \varphi' \diamondsuit\!\!\rightarrow \psi'$:
  $S, w \vDash \varphi' \diamondsuit\!\!\rightarrow \psi'$ is true, iff both a world $w' \in W_w$ exists, such that $S, w' \vDash \varphi'$ and 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 $S, w^* \vDash \varphi \wedge \psi$. And the corresponding semantic game is $\mathfrak{S}_p[\varphi' \diamondsuit\!\!\rightarrow \psi', w, S] = ((\varphi' \diamondsuit\!\!\rightarrow \psi', w)_p, E, G, \rightarrowtail)$. Every play of the semantic game of counterfactuals $\mathfrak{S}_p[\varphi' \diamondsuit\!\!\rightarrow \psi', w, S]$ begins with one

22

of the two forms of partial plays $(\varphi' \diamondsuit\!\!\rightarrow \psi', w)_p (\varphi' \diamondsuit\!\!\rightarrow \psi', w, w', r)_{np} (\varphi' \wedge \psi', w')_p$ and $(\varphi' \diamondsuit\!\!\rightarrow \psi', w)_p (\varphi' \diamondsuit\!\!\rightarrow \psi', w, w', r)_{np} (\neg\varphi', w^*)_p$ where $w, w', w^* \in W$; $r, r^* \in \mathbb{R}$; $w \overset{r}{\rightsquigarrow} w'$; $w \overset{r^*}{\rightsquigarrow} w^*$; $r^* < r$ and $np \in \{a, d\} \setminus \{p\}$ (see figure 6 moves 11,12,13).

Suppose no φ'-world is accessible from w. Then $S, w' \nvDash \varphi'$ for each world $w' \in W_w$. And given the induction assumption, the player np has a winning strategy for the game $\mathfrak{S}_p[\varphi', w', S]$. That strategy is also a winning strategy for the player np and game $\mathfrak{S}_p[\varphi' \diamondsuit\!\!\rightarrow \psi', w, S]$, if it maps $(\varphi' \diamondsuit\!\!\rightarrow \psi', w, w', r)_{np}$ to $((\varphi' \diamondsuit\!\!\rightarrow \psi', w, w', r)_{np}, (\varphi' \wedge \psi', w')_p)$ and it maps $(\varphi' \wedge \psi', w, p)_{np}$ to $((\varphi' \wedge \psi', w, p)_{np}, (\varphi', w')_p)$. It follows that the player np wins the game $\mathfrak{S}_p[\varphi' \diamondsuit\!\!\rightarrow \psi', w, S]$.

Now instead suppose that there exists at least one φ'-world that is accessible from w. Applying the limit assumption, it follows that a φ'-world accessible from w exists, such that no other φ'-world accessible from w is more similar to w than it. And if and only if such a world $w' \in W_w$ and an $r \in \mathbb{R}$ with $w \overset{r}{\rightsquigarrow} w'$ and $S, w' \vDash \varphi' \wedge \psi'$ exist, such that no world $w^* \in W_w$ and $r^* \in \mathbb{R}$ with $w \overset{r^*}{\rightsquigarrow} w^*$ and $S, w^* \vDash \varphi'$ and $r^* < r$ exist, then w' is a world, such that for every world $w'' \in W_w$ and $r'' \in \mathbb{R}$ with $w \overset{r''}{\rightsquigarrow} w''$ and $S, w'' \vDash \varphi' \wedge \neg\psi'$, it holds that $r \leq r''$ and $w \overset{r}{\rightsquigarrow} w'$ and $S, w' \vDash \varphi' \wedge \psi'$. Which means that whenever a $(\varphi' \wedge \psi')$-world w' accessible from w exists, such that there is no φ'-world w'' accessible from w that is more similar to the world w than w', it is true that $S, w \vDash \varphi' \diamondsuit\!\!\rightarrow \psi'$. We thus show that the player p wins the game $\mathfrak{S}_p[\varphi' \diamondsuit\!\!\rightarrow \psi', w, S]$ whenever there is a $(\varphi' \wedge \psi')$-world accessible from w, such that there is no φ'-world accessible from w that is more similar to w than it.

Suppose that a world $w' \in W_w$ and an $r \in \mathbb{R}$ with $w \overset{r}{\rightsquigarrow} w'$ and $S, w' \vDash \varphi' \wedge \psi'$ exist, such that no world $w^* \in W_w$ and $r^* \in \mathbb{R}$ with $w \overset{r^*}{\rightsquigarrow} w^*$ and $S, w^* \vDash \varphi'$ and $r^* < r$ exist. Given the induction assumption, the player p has a winning strategy for the game $\mathfrak{S}_p[\varphi' \wedge \psi', w', S]$. And since there is no φ'-world that is more similar to w than w', the player p also has a winning strategy for every game of the form $\mathfrak{S}_p[\neg\varphi', w^*, S]$ where $w^* \in W_w$ with $r^* \in \mathbb{R}$ and $r^* < r$. We can define a strategy in such a way—picking and choosing the mappings from the previously mentioned winning strategies—that it is a winning strategy for the player p for both of the aforementioned forms of games. If that strategy also maps $(\varphi' \diamondsuit\!\!\rightarrow \psi', w)_p$ to $((\varphi' \diamondsuit\!\!\rightarrow \psi', w)_p, (\varphi' \diamondsuit\!\!\rightarrow \psi', w, w', r)_{np})$, then it is a winning strategy for the player p and game $\mathfrak{S}_p[\varphi' \diamondsuit\!\!\rightarrow \psi', w, S]$.

Finally suppose that a world $w^* \in W_w$ and an $r^* \in \mathbb{R}$ with $w \overset{r^*}{\rightsquigarrow} w^*$ and $S, w^* \vDash \varphi'$ exist, such that no world $w' \in W_w$ and $r \in \mathbb{R}$ with $w \overset{r}{\rightsquigarrow} w'$ and $S, w' \vDash \varphi' \wedge \psi'$ and $r \leq r^*$ exist. Then, given the induction assumption, it holds for each world $w' \in W_w$ and $r \in \mathbb{R}$ with $w \overset{r}{\rightsquigarrow} w'$ that the player np has a winning strategy for either the game $\mathfrak{S}_p[\varphi' \wedge \psi', w', S]$ or $\mathfrak{S}_p[\neg\varphi', w^*, S]$. Again we can construct a strategy by combining all of those winning strategies' essential mappings, such that the constructed strategy is a winning strategy with respect to every game that one of the aforementioned strategies is a winning strategy of. If that strategy then maps $(\varphi' \diamondsuit\!\!\rightarrow \psi', w, w', r)_{np}$ to $((\varphi' \diamondsuit\!\!\rightarrow \psi', w, w', r)_{np}, (\varphi' \wedge \psi', w')_p)$ iff the constructed strategy contains a winning strategy for the player np and game $\mathfrak{S}_p[\varphi' \wedge \psi', w', S]$; and otherwise maps $(\varphi' \diamondsuit\!\!\rightarrow \psi', w, w', r)_{np}$ to $((\varphi' \diamondsuit\!\!\rightarrow \psi', w, w', r)_{np}, (\neg\varphi', w^*)_p)$ for each w', then it is also a winning strategy for the player np and game $\mathfrak{S}_p[\varphi' \diamondsuit\!\!\rightarrow \psi', w, S]$.

23

- Induction Step $\varphi = \varphi' \,\Box\!\!\rightarrow\, \psi'$:

$S, w \vDash \varphi' \,\Box\!\!\rightarrow\, \psi'$, iff either no world $w' \in W_w$ exists, such that $S, w' \vDash \varphi'$, or 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 $S, w^* \vDash \neg\varphi' \vee \psi'$. And the corresponding semantic game is $\mathfrak{S}_p[\varphi' \,\Box\!\!\rightarrow\, \psi', w, S] = ((\varphi' \,\Box\!\!\rightarrow\, \psi', w)_p, E, G, \twoheadrightarrow)$. Every play of the semantic game $\mathfrak{S}_p[\varphi' \,\Box\!\!\rightarrow\, \psi', w, S]$ begins with one of the two forms of partial plays $(\varphi' \,\Box\!\!\rightarrow\, \psi', w)_p(\varphi' \,\Box\!\!\rightarrow\, \psi', w, p)_{np}(\varphi' \,\Box\!\!\rightarrow\, \psi', w, w', r)_p(\neg\varphi' \vee \psi', w')_p$ and $(\varphi' \,\Box\!\!\rightarrow\, \psi', w)_p(\varphi' \,\Box\!\!\rightarrow\, \psi', w, p)_{np}(\varphi' \,\Box\!\!\rightarrow\, \psi', w, w', r)_p(\varphi', w^*)_p$ where $w, w', w^* \in W$; $r, r^* \in \mathbb{R}; w \overset{r}{\rightsquigarrow} w'; w \overset{r^*}{\rightsquigarrow} w^*; r^* < r$ and $np \in \{a, d\} \setminus \{p\}$ (see figure 6 moves 14,15,16,17).

Suppose no φ'-world is accessible from w. Then $S, w' \nvDash \varphi'$ for each world $w' \in W_w$. And given the induction assumption, the player p has a winning strategy each game $\mathfrak{S}_{np}[\varphi', w', S]$ with $w' \in W_w$. It follows that the player p also has a winning strategy for each game $\mathfrak{S}_p[\neg\varphi', w', S]$ with $w' \in W_w$ and moreover has a winning strategy for each game $\mathfrak{S}_p[\neg\varphi' \vee \psi', w', S]$ with $w' \in W_w$. Those strategies can in turn be combined into one strategy that is a winning strategy for the player p and every game of the form $\mathfrak{S}_p[\neg\varphi' \vee \psi', w', S]$ with $w' \in W_w$. If that strategy also maps every game state of the form $(\varphi' \,\Box\!\!\rightarrow\, \psi', w, w', r)_p$ with $w' \in W_w$; $r \in \mathbb{R}$ and $w \overset{r}{\rightsquigarrow} w'$ to $((\varphi' \,\Box\!\!\rightarrow\, \psi', w, w', r)_p, (\neg\varphi' \vee \psi', w')_p)$, then it also a winning strategy for the player p and game $\mathfrak{S}_p[\varphi' \,\Box\!\!\rightarrow\, \psi', w, S]$.

Now instead suppose that there exists at least one φ'-world that is accessible from w. Applying the limit assumption, it follows that a φ'-world accessible from w exists, such that no other φ'-world accessible from w is more similar to w than it. And if and only if it is the case that a $w' \in W_w$ and an $r \in \mathbb{R}$ with $w \overset{r}{\rightsquigarrow} w'$ and $S, w' \vDash \varphi' \wedge \neg\psi'$ exist, such that no world $w^* \in W_w$ and $r^* \in \mathbb{R}$ with $w \overset{r^*}{\rightsquigarrow} w^*$ and $S, w^* \vDash \varphi' \wedge \psi'$ and $r^* < r$ exist, then it holds that no world $w'' \in W_w$ and $r'' \in \mathbb{R}$ with $w \overset{r''}{\rightsquigarrow} w''$ and $S, w'' \vDash \varphi'$ exist, such that for each world $w^* \in W_w$ and $r^* \in \mathbb{R}$ with $w \overset{r^*}{\rightsquigarrow} w^*$ and $r^* \leq r$, it holds that $S, w^* \vDash \neg\varphi' \vee \psi'$. Which means that whenever a $(\varphi' \wedge \neg\psi')$-world w' accessible from w exists, such that there is no φ'-world w^* accessible from w that is more similar to the world w than w', it is true that $S, w \nvDash \varphi' \,\Box\!\!\rightarrow\, \psi'$. We thus show that the player p wins the game $\mathfrak{S}_p[\varphi' \,\Box\!\!\rightarrow\, \psi', w, S]$ whenever there is no $(\varphi' \wedge \neg\psi')$-world accessible from w, such that there is no φ'-world accessible from w that is more similar to w than it.

Suppose that a world $w^* \in W_w$ and an $r^* \in \mathbb{R}$ with $w \overset{r^*}{\rightsquigarrow} w^*$ and $S, w^* \vDash \varphi'$ exist, such that no world $w' \in W_w$ and $r \in \mathbb{R}$ with $w \overset{r}{\rightsquigarrow} w'$ and $S, w' \vDash \varphi' \wedge \neg\psi'$ and $r \leq r^*$ exist. Then $w^* \vDash \varphi' \wedge \psi'$ and each world at least as similar to w as w^* is also a $(\neg\varphi' \vee \psi')$-world. Given the induction assumption, the player p has a winning strategy for the game $\mathfrak{S}_p[\varphi' \wedge \psi', w^*, S]$. It follows that the player p also has a winning strategy for the game $\mathfrak{S}_p[\psi', w^*, S]$ and moreover has a winning strategy for $\mathfrak{S}_p[\neg\varphi' \vee \psi', w^*, S]$. Given the induction assumption, it also holds that the player p has a winning strategy for the game $\mathfrak{S}_p[\varphi', w^*, S]$. We can combine the latter two strategies into one strategy that is a winning strategy with respect to both the games $\mathfrak{S}_p[\neg\varphi' \vee \psi', w^*, S]$ and $\mathfrak{S}_p[\varphi', w^*, S]$. If that strategy also maps every game state of the form $(\varphi' \,\Box\!\!\rightarrow\, \psi', w, w', r)_p$ with $w' \in W_w \setminus \{w^*\}$; $r \in \mathbb{R}$ and $w \overset{r}{\rightsquigarrow} w'$ to $((\varphi' \,\Box\!\!\rightarrow\, \psi', w, w', r)_p, (\varphi', w^*)_p)$; and maps the game state $(\varphi' \,\Box\!\!\rightarrow\, \psi', w, w^*, r^*)_p$ to $((\varphi' \,\Box\!\!\rightarrow\, \psi', w, w^*, r^*)_p, (\neg\varphi' \vee \psi', w^*)_p)$, it is a winning strategy for the player p and game $\mathfrak{S}_p[\varphi' \,\Box\!\!\rightarrow\, \psi', w, S]$.

24

Suppose that a world $w' \in W_w$ and an $r \in \mathbb{R}$ with $w \overset{r}{\leadsto} w'$ and $S, w' \vDash \varphi' \wedge \neg \psi'$ exist, such that no world $w^* \in W_w$ and $r^* \in \mathbb{R}$ with $w \overset{r^*}{\leadsto} w^*$ and $S, w^* \vDash \varphi'$ and $r^* < r$ exist. Then given the induction assumption, the player np has a winning strategy for each game of the form $\mathfrak{S}_p[\varphi', w^*, S]$ where $w^* \in W_w$, $r \in \mathbb{R}$, $w \overset{r}{\leadsto} w^*$ and $r^* < r$; and also has a winning strategy for the game $\mathfrak{S}_p[\neg \varphi' \vee \psi', w', S]$. Again we can construct a strategy that is a winning strategy for the player np with respect to each game of the form $\mathfrak{S}_p[\varphi', w^*, S]$ where $w^* \in W_w$, $r \in \mathbb{R}$, $w \overset{r}{\leadsto} w^*$ and $r^* < r$; and also is a winning strategy for the player np and game $\mathfrak{S}_p[\neg \varphi' \vee \psi', w', S]$. And if that strategy maps $(\varphi' \mathbin{\Box\!\!\rightarrow} \psi', w, p)_{np}$ to $((\varphi' \mathbin{\Box\!\!\rightarrow} \psi', w, p)_{np}, (\varphi' \mathbin{\Box\!\!\rightarrow} \psi', w, w', r)_p)$, it is a winning strategy for the player np and game $\mathfrak{S}_p[\varphi' \mathbin{\Box\!\!\rightarrow} \psi', w, S]$.

$\square$

7   The interactive demonstration

This section introduces the interactive demonstration of the semantic game of counterfactuals and explores its presentation and correspondence to the semantic game. We will begin by explaining the game metaphor and offering an overview of the user interface. Afterwards, we will look in much greater detail at some parts of the web demonstration's visual presentation and explain user interaction and the process of playing the game. Then, we highlight some limitations of the web demonstration and differences to the semantic game. And finally, we discuss the algorithm that is used to compute a player's best move in a given game position.

7.1   Game metaphor

The semantic game's demonstration employs an embellishing metaphor for the course of the game in an effort to veil the complexity of the semantic game of counterfactuals and more succinctly convey its notions. The metaphor's most general sense is that the players are pilots of a spaceship that are tasked with traveling between possible worlds in order to reach a second earth. This interpretation lets Lewis' notion of the realism of possible worlds become palpable and provides an intuitive explanation for both players having the same current world, because they travel together. In order to discern the players, the metaphor uses the term pilot to describe the defender and copilot to describe the attacker, which also has a drawback. As most people tend to view the relationship between a pilot and copilot as one of collaboration and teamwork, it is not immediately obvious that the relationship between the pilot and copilot is adversarial in this case. The metaphor explains the adversity of the pilots by both of them wanting to be the first human to land on the second earth and figuratively make a small step for a man, but a giant leap for mankind. The act of landing is thus a metaphor for reaching the $\top$-symbol and winning the reachability game. The loss condition that a player that is stuck in turn is explained by that player being dismissed from their position as pilot for stranding the spaceship. Returning to our general metaphor we can explain the general makeup of a game state $(\varphi, w)_p$. The formula φ is a schedule that the pilots have to strictly abide by. This schedule was precalculated by the boardcomputer, such that it is definitely possible to reach a habitable world by adhering to it. The world w is merely the world the pilots' spaceship is currently at. And the active player p is the pilot whose shift it currently is. The pilots work in shifts determined by the schedules shift switches that are metaphors for negations. Since different atomic formulas are true at different possible worlds, we call atomic formulas habitability tests. A world can

only be landed on, if a habitability test was successfully conducted there. The possibility and necessity operators are simply jumps of the spaceship to another reachable world. And the counterfactual might and counterfactual would are explained as the opposing pilot's meddling with the board computer in order to temporarily limit the spaceships jump range.

7.2   User interface Layout

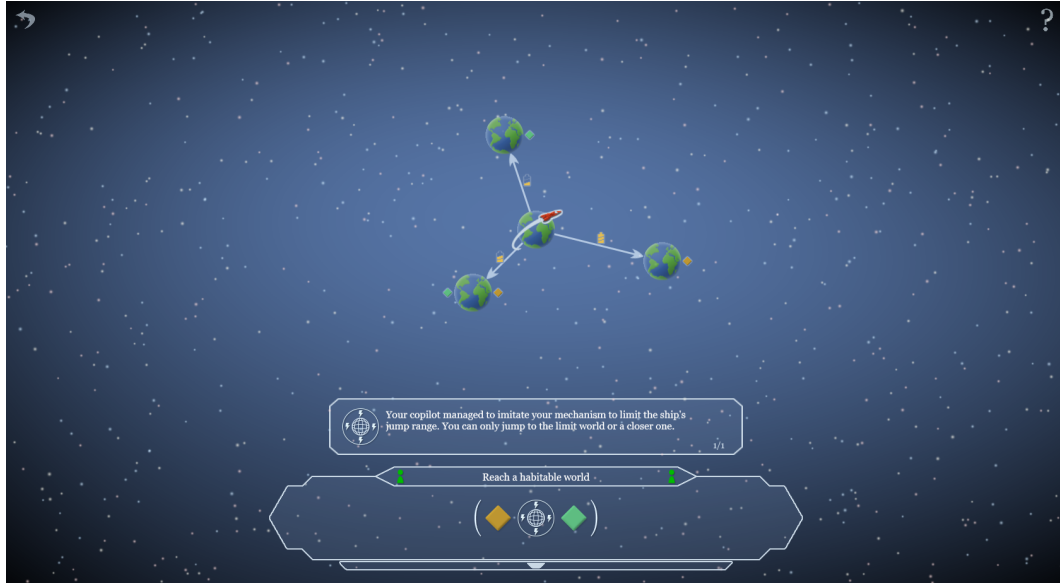The game screen of the interactive demonstration appears as follows.

Fig. 15: User interface

The center of the screen is taken up by a graph representation of the counterfactual Kripke structure. The similarity relation is represented through directed, labeled edges between possible worlds. Those edges are labeled with battery icons with 3 distinct charge states. The charge state of the battery is the value of the Kripke structure's respective measure of similarity between worlds. And for the sake of simplicity of this demonstration the similarity values are constrained to 4 values $0, 1, 2$ and 3, in ascending dissimilarity. Each world has an implicit recursive edge with a similarity value of 0.

At the bottom of the screen there is a supposition panel, where the current formula resides.

center of the screen, there is a representation of the Kripke structure the game is played on On the bottom there is the sup panel holding the current formula introduction textbox and hints in the top right corner objective line

7.3 Representing counterfactual formulas

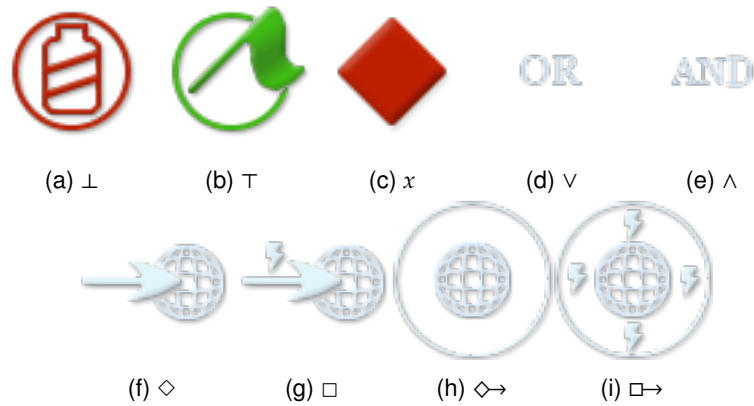(a) ⊥ (b) ⊤ (c) x (d) ∨ (e) ∧

(f) ◇ (g) □ (h) ◇→ (i) □→

Fig. 16: Icon representations

formula examples Only 26 Atomic formulas (Bc more are not necessary) Why so many operators? Screenspace. icon criticism (antecedent/consequent are not treated equally, that needs to be reflected in Cf Might,Cf Would)

7.4 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

7.5 Artificial intelligence

Algorithm 1 Adjusted MiniMax Algorithm

Input: s (game state)

```
 1: function BACKTRACKING(s)                                                      ▷ Calculate best move
 2:     Q ← [s]                                          ▷ Setup queue-, expanded node- and score lists
 3:     E ← []
 4:     SC ← []
 5:     while Q not empty do
 6:         current ← PEEK(Q)
 7:         e ← false
 8:         states ← NEXTSTATES(current)
 9:         if current not in E then                                                  ▷ Expand current
10:             PUSH(E, current)
11:             e ← true
12:             for s in states do
13:                 if s not in E then                                 ▷ Schedule s for next expansion
14:                     if s in Q then
15:                         REMOVEEACH(Q,s)
16:                     end if
17:                     PUSH(Q,state)
18:                 end if
19:             end for
20:         end if
21:         if e = false then        ▷ Process current, if it has not been expanded this iteration
22:             score ← CALCSCORE(current)
23:             PUSH(SC,score)
24:             POP(Q)                                             ▷ Remove the processed element from Q
25:         end if
26:     end while
27: end function
28: PEEK retrieves the last element of a list, without removing it from the list
29: POP retrieves the last element of a list and removes it from the list
30: PUSH appends an element to the end of a list
31: NEXTSTATES calculates the next game states and returns them
32: CALCSCORE calculates a score for the quality of a game state, based on its successor
    game states
```

$$
sc(s) = \begin{cases}
(0,0), & \text{if } s = (\top, w)_a \text{ or the defender is stuck at } s \\
(1,0), & \text{if } s = (\top, w)_d \text{ or the attacker is stuck at } s \\
(\Sigma_{i=1}^n v_i/n, \Sigma_{i=1}^n dec_i/n + 1), & \text{if the defender has } n \text{ playable moves } m_{1,...,n} \text{ with } sc(m_i) = (v_i, dec_i) \text{ at } s \\
(v_{min}, dec_{min}), & \text{if the attacker has } n \text{ playable moves } m_{1,...,n} \text{ with } sc(m_i) = (v_i, dec_i) \text{ at } s \text{ and} \\
& v_{min} \leq v_j \text{ and } dec_{min} \leq dec_j \text{ for each } j \in \{1, ..., n\})
\end{cases}
$$

$$(1)$$

The algorithm expands game state nodes like DFS, until it encounters a node that it has already expanded. In that case it processes that node and calculates our blunderscore metric for it. Since expanded nodes remain in the queue until they are processed, the queue can

be intuited like a current path of search or processing order. Expanding a node is akin to considering all its dependencies and processing them first. It may happen, that a node that is already scheduled for processing (in the queue) is encountered another time. Then that node is deleted from the queue and pushed onto it (see lines 13-18). This means, simply that it is promoted to be processed earlier. And because the game graph does not contain cyclic dependencies, this does not run into infinite promotions of the processing order. In other words, because there are no cycles in the graph of game states, the algorithm will at some point encounter a node that has no successors and then be able to process it. If edges to the already processed nodes are left out of consideration, then there always is such a leaf node to process.

Since the AI in the game only ever calculates moves for the attacker, the definition of blunderscores is skewed that way. A converse formulation is easy to derive.

8   Results

9   old section 7 layout

- explain correspondence - representation of similarity graph, formulas - motivate and explain the UI and interaction paradigm (top/bottom) - justify limit assumption/problems - explain limitations/differences of the game representation and game engine

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.