An interactive demonstration of counterfactual truth conditions[*]

Proposal for a Bachelor Thesis (v0.5, 3 August 2022)

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

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

Problem: An approachable showcase of counterfactuals

Counterfactuals are in essence conditional statements carrying information about the relationship of a false *antedecent* and a *consequent*. Their truth value answers the question: "If it were the case that A, would it then be the case that B?". *Counterfactual thought*—that is the thought of alternate outcomes—[Byr16] is common for humans to engage in and is consequently subject to a great volume of philosophical and mathematical discourse.

Considering the amount of information available and the various propositions on how to mathematically model counterfactuals [Sta68, ST70, Lew73], gaining an understanding of the nuances of counterfactual truth conditions tends to be a difficult endeavour for any layman. I propose to remedy this issue by implementing a computer game based on a semantic game of counterfactuals.

Approach: A computer game on counterfactuals

The aforementioned semantic game of counterfactuals expresses the evaluation of *Lewis's counterfactual truth conditions* [Lew73] in terms of a two-player game wherein a defender tries to prove a counterfactual sentence and an attacker attempts to disprove it.

The game state is described by a tuple holding a counterfactual sentence φ and a world w at which φ ought to be shown. Additionally the subscript of a tuple denotes whose players turn it is. Tuples with designators "vac" and "cf" are exceptions to that syntax, indicating supposed vacuous truth or truth at all worlds at least as similar to w as w'. Here are some examples:

$$(\varphi, w)_{d/a}, \;\; (vac, \varphi, w)_a, \;\; (cf, \varphi, \psi, w, w', r)_a$$

[*] Further title proposals: (B) Creating an educational computer game about counterfactuals in terms of a centered system of spheres (C) Implementing a computer game illustrating the truth conditions of counterfactuals as variably strict conditionals

The constituents of a counterfactual sentence φ are negations $\neg$, disjunctions $\vee$, the counterfactual would operator $\Box\!\!\rightarrow$, the bottom symbol $\bot$ and atomic statements. To be able to resolve atomic statements each world w has a set of true atomic statements $F(w)$ associated with it. Moreover a ternary relation $\rightsquigarrow$ expresses similarity between worlds from the standpoint of one world.

Given those remarks, the concrete rules of the game are as stated in figure 3.

Approach: Visual representation

I intend to create a software solution that is intuitive and easy to use. The user is meant to be able to learn about counterfactuals, with little prior knowledge or time investment. Hence a minimalistic and functional user interface has to be developed.

For the purposes of visual clarity, usability and simplicity the application will provide finite selections of exemplar worlds as stand-ins for potentially infinite sets of accessible worlds and visualize them as nodes of a graph as shown in figure 1.

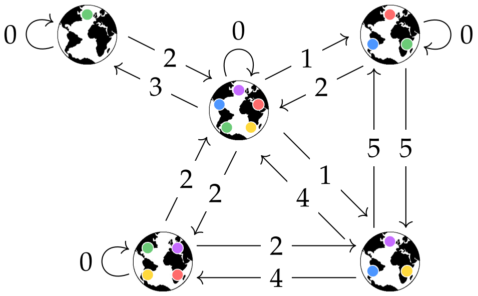

Fig. 1: similarity graph of exemplar worlds

The directed edges of the graph express dissimilarity from the standpoint of one world to another. And the colored dots represent atomic statements from the set $F(w)$, that each world has associated with it.

Additionally I will determine a visual representation of counterfactual sentences from amongst various possibilities such as natural language, logical expressions and logical expression trees. While natural language is the most intuitive representation, it lacks the unambiguity of logical constructs and may create long unreadable sentences. Logical expressions, however, tend to be difficult to read and comprehend for laymen. Here are some examples of the same counterfactual statement expressed in different ways in (1), (2) and figure 2.

"If Alexander the Great had not died at the age of 32 and attacked europe, the Romans would have defeated him" (1)

$$\neg(\varphi_1 \vee \neg\varphi_2) \mathrel{\Box\!\!\rightarrow} \psi \tag{2}$$

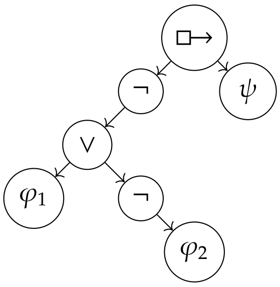

Fig. 2: counterfactual sentence as a logical expression tree

Method: Deployment as a lightweight web application

In order to deliver an easily accessible software solution I intend to make use of the Phaser 3 Framework for Typescript to develop a web-application. Phaser 3 is well suited for this task because it is a lightweight framework, that fits the scope of the project and allows for distribution of the software with a small barrier to entry.

Definitions

0.1 Counterfactual formulas

$Atoms = \{x, y, ...\}$
$\Phi = \{\varphi, \psi, ...\}$
$\varphi, \psi ::= \bot \mid x \mid \neg\varphi \mid \varphi \vee \psi \mid \varphi \;\Box\!\!\rightarrow\; \psi$

0.2 Worlds

$W = \{w, v, ...\}$

0.3 Facts

$F \colon W \rightarrow 2^{Atoms}$

0.4 Similarity relation

$\rightsquigarrow \colon W \times \mathbb{R} \times W$

0.5 Accessible worlds

$W_w = \{w' \mid w \overset{r}{\rightsquigarrow} w'\}$

0.6 Truth conditions of counterfactual logic

$w \models \bot$ is always false.
$w \models x$ iff $x \in V(w)$.
$w \models \varphi \vee \psi$ iff ($w \models \varphi$ or $w \models \psi$)

$w \vDash \varphi \mathrel{\Box\!\!\rightarrow} \psi$, if no world w' and r exist, such that $w' \vDash \varphi$ and $w \overset{r}{\rightsquigarrow} w'$.

$w \vDash \varphi \mathrel{\Box\!\!\rightarrow} \psi$, if a world w' and an r exist, such that $w' \vDash \varphi$ and $w \overset{r}{\rightsquigarrow} w'$ and for each world $w*$, for which a $r* \leq r$ exists, such that $w \overset{r*}{\rightsquigarrow} w*$, $w* \vDash \psi \vee \neg\varphi$ holds true.

0.7 Similarity graph

$$G = (V, E, F), \text{ such that } V \subseteq W \text{ and } E \subseteq \rightsquigarrow \tag{3}$$

Rules of the semantic game

$$(\bot, w)_a \quad \text{Attacker wins} \tag{4}$$

$$(\neg\bot, w)_a \quad \text{Defender wins} \tag{5}$$

$$(x, w)_a \xrightarrow{x \in F(w)} (\neg\bot, w)_{d/a} \tag{6}$$

$$(x, w)_a \xrightarrow{x \notin F(w)} (\bot, w)_{d/a} \tag{7}$$

$$(\neg x, w)_a \xrightarrow{x \in F(w)} (\bot, w)_{d/a} \tag{8}$$

$$(\neg x, w)_a \xrightarrow{x \notin F(w)} (\neg\bot, w)_{d/a} \tag{9}$$

$$(\neg\neg\varphi, w)_a \rightarrow (\varphi, w)_{d/a} \tag{10}$$

$$(\varphi \vee \psi, w)_d \rightarrow (\varphi, w)_{d/a} \tag{11}$$

$$(\varphi \vee \psi, w)_d \rightarrow (\psi, w)_{d/a} \tag{12}$$

$$(\neg(\varphi \vee \psi), w)_a \rightarrow (\neg\varphi, w)_{d/a} \tag{13}$$

$$(\neg(\varphi \vee \psi), w)_a \rightarrow (\neg\psi, w)_{d/a} \tag{14}$$

$$(\varphi \mathrel{\Box\!\!\rightarrow} \psi, w)_d \rightarrow (vac, \varphi, w)_a \tag{15}$$

$$(vac, \varphi, w)_a \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\neg\varphi, w')_{d/a} \tag{16}$$

$$(\varphi \mathrel{\Box\!\!\rightarrow} \psi, w)_d \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (cf, \varphi, \psi, w, w', r)_a \tag{17}$$

$$(cf, \varphi, \psi, w, w', r)_a \rightarrow (\varphi, w')_{d/a} \tag{18}$$

$$(cf, \varphi, \psi, w, w', r)_a \xrightarrow{[w \overset{r^*}{\rightsquigarrow} w^*, r^* \leq r]} (\neg\varphi \vee \psi, w^*)_d \tag{19}$$

$$(\neg(\varphi \mathrel{\Box\!\!\rightarrow} \psi), w)_a \rightarrow (vac, \varphi, w)_d \tag{20}$$

$$(vac, \varphi, w)_d \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (\varphi, w')_{d/a} \tag{21}$$

$$(\neg(\varphi \mathrel{\Box\!\!\rightarrow} \psi), w)_a \xrightarrow{[w \overset{r}{\rightsquigarrow} w']} (cf, \varphi, \psi, w, w', r)_d \tag{22}$$

$$(cf, \varphi, \psi, w, w', r)_d \rightarrow (\neg\varphi, w')_{d/a} \tag{23}$$

$$(cf, \varphi, \psi, w, w', r)_d \xrightarrow{[w \overset{r^*}{\rightsquigarrow} w^*, r^* \leq r]} (\neg(\neg\varphi \vee \psi), w^*)_a \tag{24}$$

Fig. 3: rules of the semantic game of counterfactuals

Schedule

In accordance with StuPO Bachelor Informatik 2015, the elaboration of the thesis is scheduled over a period of 20 weeks:

3 weeks Familiarization with theory & Game design.

1 week Implementation of a semantic game engine. (Output: Software)

3 weeks Implementation of the user interface. (Output: Counterfactual representation, Graph representation, Menu)

2 weeks Creation of levels or levelgenerator. (Output: Software)

2 weeks Testing & QA.

6 weeks Write thesis. (Output: Thesis)

3 weeks Puffer

Thesis Outline (Sketch)

1. Introduction
   (a) Problem description
   (b) Context / related work
   (c) Overview
2. Formulation of the semantic game of counterfactuals
   (a) Definitions (worlds, facts at worlds, graph of worlds)
   (b) Game rules
3. Implementation
   (a) Design decisions (visual representation of cfs, finite graph / exemplar worlds)
   (b) Discussion of the framework
   (c) Implementation of the semantic game engine
   (d) Evaluation of the implementation
4. Conclusion
   (a) Summary
   (b) Outlook

References

[Byr16] Ruth M.J. Byrne. Counterfactual thought. *Annual Review of Psychology*, 67(1):135–157, 2016. PMID: 26393873. `arXiv:https://doi.org/10.1146/annurev-psych-122414-033249`, `doi:10.1146/annurev-psych-122414-033249`.

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

[ST70] Robert C. Stalnaker and Richmond H. Thomason. *A Semantic Analysis of Conditional Logic*. 1970.

[Sta68] Robert C. Stalnaker. *A Theory of Conditionals*. Basil Blackwell, 1968.