Towards a Modal Logic of Causal Counterfactuals

Kenneth Lai and **James Pustejovsky**
Department of Computer Science
Brandeis University
Waltham, MA 02453
{klai12, jamesp}@brandeis.edu

Abstract

Previous work has studied the relationship between a causal modeling semantics for counterfactual sentences, as encoded in structural equations, and possible-world approaches common in linguistics and philosophy. However, such comparisons have generally not dealt with the possible-world models themselves, but rather with the underlying logic. Furthermore, previous work has not considered the role of context, which is crucial to interpreting the meaning of counterfactuals in natural language. We present a possible-world semantics and a modal logic for context-dependent causal counterfactuals, inspired by logics for strategic reasoning.

Introduction

In order for artificial intelligence systems to reason effectively about the world, generalizing to new situations much like a human would, counterfactual reasoning is essential (Pearl and Mackenzie 2018). If it were the case that A, would it have been the case that B? Or, to give a classic example from Lewis (1973), if kangaroos had no tails, would they topple over? Much work has been done on counterfactual reasoning, in computer science as well as in other disciplines.

One idea, common in linguistics and philosophy, is to define counterfactuals in terms of possible worlds (Lewis 1973; Stalnaker 1968; Kratzer 1981). In the actual world, kangaroos have tails, but we can think of a possible world in which they do not, and consider whether they topple over in that world (let us assume that they do). Of course, we can think of many such worlds; for example, in a world where kangaroos had no tails but used crutches, perhaps they would not topple over. Alternatively, we can think of a world where the laws of physics were changed such that tail-less kangaroos would remain upright. Crucially, we only consider the closest worlds to the actual world, according to some distance metric, or ordering of worlds.

Formally, this is done by setting the possible worlds in a Kripke structure $\langle W, R, L \rangle$, where W is the set of worlds, R is the accessibility relation between worlds, and L is the labeling function that maps worlds to the sets of propositions true at those worlds. Comparative similarity between worlds is then encoded in the accessibility relation R: for two worlds w and w', $R(w, w')$ if and only if w' is sufficiently similar to w.

This raises the question of what distance metric should be used. A key insight from Pearl (2000) is that counterfactuals rely on the notion of cause and effect. Specifically, the distance metric should be consistent with the causal laws in effect in the actual world: worlds that differ in their causal laws are more distant than worlds whose laws are the same.

In Pearl's theory, causal laws are expressed in terms of structural equations. An equation $a = f(b)$ denotes that the value of a is dependent on the value of b. We can then reason about what the value of a would have been, if the value of b had been different. The set of structural equations, together with sets of endogenous and exogenous variables, define a causal model.

Pearl (2000) showed that for nonrecursive systems, the axioms of composition and effectiveness are sufficient to derive the axioms of Lewis' logic, and vice versa. Nevertheless, there remain differences in the kinds of counterfactual sentences expressible in each theory. Pearl identifies possible worlds with instantiations of variables in a causal model. This is sufficient to express sentences of the form $\mathbf{A} \mathrel{\Box\!\!\rightarrow} \mathbf{B}$, where $\mathrel{\Box\!\!\rightarrow}$ is Lewis' counterfactual operator, and $\mathbf{A}$ and $\mathbf{B}$ are conjunctions of variable values. However, more complex sentences, with arbitrary antecedents and consequents, cannot be modeled within this framework. Some work has been done to extend the causal modeling approach to different types of counterfactuals. For example, Briggs (2012) models counterfactual sentences with counterfactual consequents (by making successive interventions on a causal model), and Boolean (non-counterfactual) antecedents (using the concept of truthmaking).

However, a possible-world semantics retains other advantages. von Fintel (2001) provides evidence that the meaning of a counterfactual depends on, and affects, the context in which it is uttered. In this case, the relevant context is the accessibility relation R. Briefly, when evaluating a counterfactual sentence $A \mathrel{\Box\!\!\rightarrow} B$, the accessibility relation is modified such that some worlds where A is true become accessible from the actual world. Crucially, after the counterfactual has been evaluated, the accessibility relation does not revert to its previous state. Once we have introduced worlds in which kangaroos use crutches, we cannot subsequently forget about them when thinking of worlds where they have no tails.

We describe a framework for counterfactual reasoning, using a possible-world semantics, that incorporates causal modeling and a role for context. We present our models in terms of concurrent game structures, an extension of Kripke structures introduced by Alur, Henzinger, and Kupferman (2002) for Alternating-time Temporal Logic (ATL), a temporal logic for multiplayer games. In this setting, we furthermore define a modal logic for causal counterfactuals, based on ATL with Intentions (ATL+I), which allows for strategic reasoning (Jamroga, van der Hoek, and Wooldridge 2005).

Causal models as concurrent game structures

In a concurrent game structure, along with a set of worlds, there also exists a set of players, each with a set of possible moves at each world. At a given world, the moves made by each player determine the transition taken to the next world. We present a formal definition of a concurrent game structure below, lightly edited from Alur, Henzinger, and Kupferman's original version (2002):

Definition 1. A *concurrent game structure* is a tuple $S = \langle A, W, P, L, D, \delta \rangle$ with the following components:

- A set A of *players*.
- A set W of *worlds*.
- A set P of *propositions*.
- A *labeling function* L. For each world $w \in W$, $L(w) \subseteq P$ is the set of propositions true at w.
- A *move function* D. For each player $a \in A$ and each world $w \in W$, let $d_a(w)$ be the set of moves (Alur, Henzinger, and Kupferman use natural numbers) available at world w to player a. For each world w, a move vector at w is a tuple $\langle m_a, ... \rangle$ of moves m_a, one for each player a, such that $m_a \in d_a(w)$. Then for each world w, $D(w) = \prod_a d_a(w)$ is the set of move vectors at w.
- A *transition function* δ. For each world $w \in W$ and each move vector $\langle m_a, ..., \rangle \in D(w)$, $\delta(w, m_a, ...) \in W$ is the world that results from world w if every player $a \in A$ chooses move m_a.

We also define a *strategy* $\sigma_a(w)$ of player a as in Jamroga, van der Hoek, and Wooldridge (2005), as a function mapping each world w to a non-empty subset of $d_a(w)$. In ATL+I, a player's strategy defines their intentions; i.e., at each world, a player will only choose a move consistent with their strategy at that world. In this way, the allowable transitions from a world depend on the strategies employed by each player.

To understand the relationship between concurrent game structures and causal models, we return to the idea that the notion of comparative similarity between worlds is encoded in the accessibility relation R. We note that the above definition refers to a transition function, rather than an accessibility relation. This is because while the transition function is fixed, the accessibility relation, i.e., the relation that defines the worlds between which transitions are allowed, is dependent at any given time on the strategies currently in force.

We also return to Pearl's idea that the distance metric depends on a notion of cause and effect (2000). This gives us an idea of what the strategies in a concurrent game structure correspond to in a causal model: structural equations.

Example: cat and vase

As an illustrative example, consider a table, on which stand a cat and a vase. Our cat is a good cat, and does not push the vase off the table in the actual world; the vase does not break. However, we want to think about what would have happened had the cat pushed the vase off the table. Let c be the proposition "the cat pushed the vase off the table", and v be "the vase broke".

The laws of physics determine the fate of the vase, based on the force applied by the cat, the distance fallen, etc. For simplicity, we will associate propositional atoms to Boolean variables, and wrap all of these effects in the single structural equation $v = c$; the vase breaking depends causally on the cat pushing the vase off the table.

Let us define a concurrent game structure for our scenario:

- $A = \{C, V\}$. For the variables c and v, we associate players C and V, respectively. Their strategies will control which worlds are considered accessible, based on the values of their associated propositions in those worlds.
- $W = \{w_{00}, w_{01}, w_{10}, w_{11}\}$. We have one world for each possible valuation of the variables c and v. For example, w_{01} is the world where c is false but v is true. The actual world in our scenario is w_{00}.
- $P = \{c, v\}$.
- $L(w_{00}) = \varnothing$, $L(w_{01}) = \{v\}$, $L(w_{10}) = \{c\}$, and $L(w_{11}) = \{c, v\}$.
- For each world $w \in W$, $d_C(w) = \{0_C, 1_C\}$ and $d_V(w) = \{0_V, 1_V, c_V, \neg c_V\}$. For a player a with associated proposition p, the move 0_a (1_a) sets p false (true) at the next state. The moves c_V and $\neg c_V$ set the value of v equal or not equal to the value of c at the next state, respectively; the existence of these moves stems from the dependence of v on c in our causal model.
- For each world $w \in W$:
    - ○ $\delta(w, 0_C, 0_V) = w_{00}$
    - ○ $\delta(w, 0_C, 1_V) = w_{01}$
    - ○ $\delta(w, 0_C, c_V) = w_{00}$
    - ○ $\delta(w, 0_C, \neg c_V) = w_{01}$
    - ○ $\delta(w, 1_C, 0_V) = w_{10}$
    - ○ $\delta(w, 1_C, 1_V) = w_{11}$
    - ○ $\delta(w, 1_C, c_V) = w_{11}$
    - ○ $\delta(w, 1_C, \neg c_V) = w_{10}$

Although there are many different possible strategies for each player, we will only be interested in a few of them. Let $\mathbf{m}_a$ denote the strategy for player a that maps each world w to move m_a, i.e. $\mathbf{m}_a(w) = \{m_a\}$. In this scenario, for the structural equation $v = c$, player V has strategy $\mathbf{c}_V$.

As for player C, because c is an exogenous variable (whose value does not depend on any other variables), there is no structural equation in the causal model. However, we

do not want to say that C has no strategy. As previously mentioned, when evaluating a counterfactual sentence, we only want to consider those worlds that are closest to the actual world. But having no strategy means placing no restrictions on which worlds are accessible from the actual world. Intuitively, given a world with some value of c, worlds with the same value of c can be considered closer to that world than worlds with the opposite value, all else being equal. Therefore, one possible strategy is to keep the value of c the same; call this strategy $\mathbf{def}_C$. Then $\mathbf{def}_C(w_{00}) = \mathbf{def}_C(w_{01}) = \{0_C\}$, and
$\mathbf{def}_C(w_{10}) = \mathbf{def}_C(w_{11}) = \{1_C\}$.

Finally, we need to define some mechanism for applying an intervention to a causal model, i.e., for a player to change their strategy. Let Σ be the set of currently active strategies, and let σ_a be a new strategy for player a. Then we can define the revise function as follows:

$$\mathit{revise}(\Sigma, \sigma_a) = \{\sigma_b | \sigma_b \in \Sigma, b \neq a\} \cup \{\sigma_{a,old} \cup \sigma_a\}$$

In essence, the revise function edits the set of active strategies, adding the moves from σ_a to the previous strategy of a, while leaving other players' strategies unchanged. This ensures that after revision, there is some accessible world where the causal intervention holds.

A modal logic for counterfactual reasoning

Now we can describe our modal logic of counterfactuals. First, from the set of active strategies, it is possible to define active transitions. A transition $(w, m_a, ...)$ is active iff for all moves m_a, and active strategies $\sigma_a \in \Sigma$, it is the case that $m_a \in \sigma_a(w)$. In our scenario, with $\Sigma = \{\mathbf{def}_C, \mathbf{c}_V\}$, the active transitions are $(w_{00}, 0_C, c_V)$, $(w_{01}, 0_C, c_V)$, $(w_{10}, 1_C, c_V)$, and $(w_{11}, 1_C, c_V)$.

Next, we define the accessibility relation R over our model. If w and w' are worlds, then $R(w, w')$ iff there exist moves $m_a, ...$ such that $(w, m_a, ...)$ is active and $\delta(w, m_a, ...) = w'$. In our scenario, $R(w_{00}, w_{00})$, $R(w_{01}, w_{00})$, $R(w_{10}, w_{11})$, and $R(w_{11}, w_{11})$.

Finally, we define the syntax and semantics of our logic. All formulas of basic modal logic (Huth and Ryan 2004) are formulas of our logic, as well as formulas of the form $(\mathbf{str}_a \sigma_a \phi)$, taken from ATL+I (Jamroga, van der Hoek, and Wooldridge 2005).

Let p be a propositional atom, ϕ be a formula, and a be a player. Then the well-formed formulas of our logic are:

$$\begin{aligned}\phi ::= &\bot | \top | p | (\neg\phi) | (\phi \wedge \phi) | (\phi \vee \phi) | (\phi \to \phi) | (\phi \leftrightarrow \phi) \\ &| (\Box\phi) | (\Diamond\phi) | (\mathbf{str}_a \sigma_a \phi)\end{aligned}$$

We define the semantics of our logic in terms of the satisfaction relation $\models$. If $\mathcal{M}$ is a concurrent game structure, Σ is a set of active strategies, and w is a world, then:

- $\mathcal{M}, \Sigma, w \models \top$
- $\mathcal{M}, \Sigma, w \not\models \bot$
- $\mathcal{M}, \Sigma, w \models p$ iff $p \in L(w)$
- $\mathcal{M}, \Sigma, w \models \neg\phi$ iff $\mathcal{M}, \Sigma, w \not\models \phi$
- $\mathcal{M}, \Sigma, w \models \phi \wedge \psi$ iff $\mathcal{M}, \Sigma, w \models \phi$ and $\mathcal{M}, \Sigma, w \models \psi$
- $\mathcal{M}, \Sigma, w \models \phi \vee \psi$ iff $\mathcal{M}, \Sigma, w \models \phi$ or $\mathcal{M}, \Sigma, w \models \psi$
- $\mathcal{M}, \Sigma, w \models \phi \to \psi$ iff $\mathcal{M}, \Sigma, w \models \psi$ whenever $\mathcal{M}, \Sigma, w \models \phi$
- $\mathcal{M}, \Sigma, w \models \phi \leftrightarrow \psi$ iff ($\mathcal{M}, \Sigma, w \models \phi$ iff $\mathcal{M}, \Sigma, w \models \psi$)
- $\mathcal{M}, \Sigma, w \models \Box\phi$ iff for each $y \in W$ with $R(w, y)$ we have $\mathcal{M}, \Sigma, y \models \phi$
- $\mathcal{M}, \Sigma, w \models \Diamond\phi$ iff there is a $y \in W$ such that $R(w, y)$ and $\mathcal{M}, \Sigma, y \models \phi$
- $\mathcal{M}, \Sigma, w \models (\mathbf{str}_a \sigma_a \phi)$ iff $(\mathcal{M}, \mathit{revise}(\Sigma, \sigma_a), w \models \phi)$.

We can now express the sentence "If the cat had pushed the vase off the table, the vase would have broken". Under the causal modeling approach, we intervene in the model to set $c = 1$. This corresponds to a strategy for C to go to a world where c is true, i.e., $\mathbf{1}_C$. Then, following von Fintel (2001), we check whether in all accessible worlds where c is true, v is also true; this is the strict conditional $\Box(c \to v)$. Therefore, the formula we want to evaluate is $\mathbf{str}_C \mathbf{1}_C \Box(c \to v)$.

To evaluate this formula at the actual world w_{00}, we revise the set of active strategies $\Sigma = \{\mathbf{def}_C, \mathbf{c}_V\}$ with the new strategy $\mathbf{1}_C$: the result is $\Sigma' = \{\mathbf{def}_C \cup \mathbf{1}_C, \mathbf{c}_V\}$. The updated strategy $\mathbf{def}_C \cup \mathbf{1}_C$ contains moves from both strategies $\mathbf{def}_C$ and $\mathbf{1}_C$; in particular, at world w_{00}, the set of moves becomes $\{0_C, 1_C\}$. This activates the transition $(w_{00}, 1_C, c_V)$, and therefore the accessibility relation expands to include $R(w_{00}, w_{11})$ (as well as $R(w_{00}, w_{00})$).

Now we test whether $\Box(c \to v)$ is true at w_{00} using the updated accessibility relation. We see whether $c \to v$ is true in all of the accessible worlds from w_{00}, namely w_{00} and w_{11}. At w_{00}, c is false, so the conditional statement is automatically true. At w_{11}, c is true, as is v, so the conditional is true in this world as well. Therefore, we can deduce that $\mathcal{M}, \Sigma, w_{00} \models \mathbf{str}_C \mathbf{1}_C \Box(c \to v)$: if the cat had pushed the vase off the table, the vase would have broken.

References

Alur, R.; Henzinger, T. A.; and Kupferman, O. 2002. Alternating-time temporal logic. *Journal of the ACM* 49(5):672–713.

Briggs, R. 2012. Interventionist counterfactuals. *Philosophical Studies* 160(1):139–166.

Huth, M., and Ryan, M. 2004. *Logic in Computer Science: Modelling and Reasoning about Systems*. Cambridge University Press.

Jamroga, W.; van der Hoek, W.; and Wooldridge, M. 2005. Intentions and strategies in game-like scenarios. In *Portuguese Conference on Artificial Intelligence*, 512–523. Springer.

Kratzer, A. 1981. Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic* 10(2):201–216.

Lewis, D. 1973. *Counterfactuals*. Blackwell.

Pearl, J., and Mackenzie, D. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Pearl, J. 2000. *Causality*. Cambridge University Press.

Stalnaker, R. C. 1968. *A Theory of Conditionals*. Dordrecht: Springer Netherlands. 41–55.

von Fintel, K. 2001. Counterfactuals in a dynamic context. *Current Studies in Linguistics Series* 36:123–152.