A Dynamic Semantics for Causal Counterfactuals

Kenneth Lai
Department of Computer Science
Brandeis University
Waltham, MA 02453
klai12@brandeis.edu

James Pustejovsky
Department of Computer Science
Brandeis University
Waltham, MA 02453
jamesp@brandeis.edu

Abstract

Under the standard approach to counterfactuals, to determine the meaning of a counterfactual sentence, we consider the "closest" possible world(s) where the antecedent is true, and evaluate the consequent. Building on the standard approach, some researchers have found that the set of worlds to be considered is dependent on context; it evolves with the discourse. Others have focused on how to define the "distance" between possible worlds, using ideas from causal modeling. This paper integrates the two ideas. We present a semantics for counterfactuals that uses a distance measure based on causal laws, that can also change over time. We show how our semantics can be implemented in the Haskell programming language.

1 Introduction and background

The problem of modeling counterfactual statements and situations has drawn much attention, in computer science, linguistics, and other disciplines. In addition to its intrinsic interest, counterfactual reasoning is important for artificial intelligence systems to be able to handle novel situations (Pearl and Mackenzie, 2018).

The classic approach to counterfactuals in linguistics and philosophy is based on a possible-worlds semantics (Lewis, 1973; Stalnaker, 1968; Kratzer, 1981). To evaluate a counterfactual, we examine a possible world where the antecedent is true, and evaluate the consequent. For example, let us consider the following classic example from Lewis (1973):

(1) If kangaroos had no tails, they would topple over.

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. However, not all possible worlds should be considered. We can consider a world in which kangaroos have no tails, but use crutches, and perhaps they would not topple over in that world. But in the actual world, kangaroos do not use crutches, so why should we consider those worlds in which they do? We therefore only consider the "closest" possible worlds to the actual world, according to some distance metric or ordering of worlds.

Formally, we have an accessibility relation R, such that $R(w, w')$ is true if and only if w' is sufficiently similar to w. This defines for each world w a context, or *modal horizon*, consisting of those worlds w' such that $R(w, w')$ (von Fintel, 2001). A counterfactual $\phi > \psi$ is true in a world if and only if in all the worlds in the modal horizon where ϕ is true, ψ is true.

Von Fintel (2001) provides evidence that this context changes over time, by considering sequences of counterfactuals. Briefly, if there are no worlds in the modal horizon where the antecedent ϕ is true, the modal horizon expands until it includes those ϕ-worlds most similar to the current world. However, after the counterfactual has been evaluated, the accessibility relation does not revert to its previous state. For example, consider the following sequence of counterfactuals (a *Lewis-Sobel sequence*):

(2) If kangaroos had no tails, they would topple over.
If kangaroos had no tails but used crutches, they would not topple over.

In the closest possible worlds in which kangaroos have no tails, they do not use crutches, and do topple over. However, in the closest worlds in which kangaroos both have no tails and use crutches, they do not topple over. The above sequence makes sense. But the next sequence of counterfactuals, with the order of the sentences reversed (a *reverse Sobel sequence*), is semantically infelicitous:

(3) If kangaroos had no tails but used crutches, they would not topple over.
#If kangaroos had no tails, they would topple over.

The first sentence expands the modal horizon to include worlds in which kangaroos have no tails and use crutches. 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. Therefore, when evaluating the second sentence, we must consider all worlds in the modal horizon where kangaroos have no tails, including both worlds in which they do and those in which they do not use crutches. In some of these worlds, they topple over, and in others, they do not.

In the classic possible-worlds approach to counterfactuals, the notion of distance or similarity between worlds is deliberately left underspecified. However, a computational implementation of counterfactuals must specify the distance metric to be used. Let us consider a possible world to be characterized by the "facts" true in that world (Kratzer, 1981). Given two worlds that differ from the actual world in the same number of facts, which one is closer? Pollock (1976) suggests that "subjunctive generalizations" are more important than other facts, while Kratzer (1981) suggests that certain facts should be "lumped" together. For example, if one looks in a mirror, one would expect to see their reflection, even if it is not currently visible (because they are not currently looking in the mirror). In other words, the facts "one looks in a mirror" and "one sees their reflection" should be lumped together: if the truth value of one fact changes, the truth of the other should change as well.

A related idea from Pearl (2000) is that the distances between worlds rely on the notion of cause and effect. Specifically, worlds that differ in their causal laws are more distant than worlds whose laws are the same. If we say that looking in a mirror causes one to see their reflection, then among worlds where one looks in the mirror, those in which they see their reflection are closer to the actual world than those where they do not.

Pearl formulates causal laws in terms of structural equations. An equation $a = f(b)$ denotes that, in a particular world, the value of a is dependent on the value of b. This allows us to 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 an enumeration of the variables, defines a causal model. While Pearl's framework cannot model all possible counterfactual sentences, others have extended the causal modeling approach to different types of counterfactuals (Briggs, 2012).

Causal modeling approaches to counterfactuals make use of interventions: changes in the causal model (Pearl, 2000). Specifically, to evaluate a counterfactual sentence, change the underlying model to make the antecedent true, and allow the change to propagate through the model. Then evaluate the consequent with respect to the new model. Briggs (2012), making connections between causal modeling and possible-worlds approaches, identifies causal models with possible worlds. Applying an intervention then corresponds to selecting the closest possible world where the antecedent is true.

In this paper, we present a semantics for counterfactual sentences that integrates causal reasoning with a dynamic semantics, such as that of Groenendijk and Stokhof (1991). Causal reasoning allows us to give an exact specification of the vague notion of "distance" between worlds, while a dynamic semantics allows us to analyze how the meaning of counterfactuals changes with context. The key idea connecting these two approaches is that causal laws can be encoded in an accessibility relation, and therefore a change in context is equivalent to an intervention in the causal model. We can formalize this using ideas from Alternating-time Temporal Logic with Intentions (ATL+I), a logic for strategic reasoning (Jamroga et al., 2005). We also present a computational implementation of our semantics in the Haskell programming language, available at `https://github.com/klai12/dscc`.

2 Causal models and concurrent game structures

Our implementation is based on *concurrent game structures*, introduced by Alur et al. (2002) as an extension of Kripke structures to open (multi-agent) systems. A Kripke structure contains a set of possible worlds, a set of propositions, and a labeling function from worlds to sets of propositions true in those worlds (Kripke, 1963). Concurrent game structures add a set of players, where each player has, for each possible world, a non-empty set of moves available at that world. The transitions available from some world are determined by the moves taken by each player at that world.

We can formally assign types to the above components as follows. We take worlds, propositions, players, and moves to be primitive types `World`, `Prop`, `Player`, and `Move`, respectively. It will be convenient to also define a type `Vector` for move vectors, i.e., which move is taken by each player, as `[(Player, Move)]`. A concurrent game structure then consists of the following six components:

- A set A of players, of type `[Player]`;
- A set W of worlds, of type `[World]`;
- A set P of propositions, of type `[Prop]`;
- A labeling function L, of type `World -> [Prop]`;
- A move function D, of type `Player -> World -> [Move]`;
- A transition function δ, of type `World -> Vector -> World`.

We now introduce the notion of a *strategy*. We adopt the definition in (Jamroga et al., 2005), as a function that, for a given player, maps each world to a non-empty subset of the moves available to that player at that world. Strategies therefore have type `World -> [Move]`. We can then define a "strategy function" σ as a non-empty subset of the move function, with type `Player -> Strategy` (or equivalently `Player -> World -> [Move]`), that specifies a strategy for each player. In ATL+I, because the strategies employed by each player restrict the set of moves from which the player will choose, and the transitions allowed from a world depend on the moves made by each player, the strategy function therefore determines which transitions are allowed. The set of allowed transitions, in turn, forms an accessibility relation that depends on the strategies used by each player.

To return to the setting of counterfactuals, we recall that in a dynamic semantics, the accessibility relation (or modal horizon) changes over time. Furthermore, using a causal modeling approach, the change in the accessibility relation is determined by an intervention in a causal model. Our proposal is to identify variables in a causal model with players in a concurrent game structure. Then we can use the strategy for a player to encode the structural equation for that variable, such that a change in strategy corresponds to an intervention in the causal model.

2.1 Example: Kangaroos, tails, and crutches

As an illustrative example, we will again consider the case of the kangaroos. Let us assume that kangaroos will topple over if and only if they have no tails and they do not use crutches; otherwise they will stay upright. Let Q, R, and S be Boolean variables corresponding to whether kangaroos have tails, use crutches, and topple over, respectively. Then we can write the structural equation $S = \neg Q \wedge \neg R$ to encode this causal law.

Now we can represent our scenario as a concurrent game structure. First, the set of players in our model is $A = \{Q, R, S\}$. Each variable in the causal model is a player in the concurrent game structure. Note that despite the use of the term "player", the players in our model are not agents, or even entities, for that matter; there are no players corresponding to "kangaroos", "tails", or "crutches".

Next we consider the space of possible worlds. We will introduce a possible world for each possible combination of moves the players can make. We will discuss the meanings of the different moves each player can make below; for now, we will say that players Q and R have two moves each (which we will call 0 and 1), and S has three moves (which we will call 0, 1, and x). Therefore, there are $2 \times 2 \times 3 = 12$ possible worlds in our concurrent game structure. We will also say that each player has the same set of available moves at each world; i.e., for all worlds w, the move function D is specified by $D(Q, w) = D(R, w) = \{0, 1\}$, and $D(S, w) = \{0, 1, x\}$. We will label the possible worlds according to the moves made by each player to arrive at that world; e.g., w_{10x} is the possible world that results when Q makes move 1, R makes move 0, and S makes move x. The combination $\{(Q, 1), (R, 0), (S, x)\}$ is then a move vector, and therefore we know that for all worlds w, the transition function $\delta(w, \{(Q, 1), (R, 0), (S, x)\}) = w_{10x}$. We can calculate the other values of the transition function in the same way.

We have specified the possible moves for each player at each world, but what do the moves mean? Although our players are not agents in the conventional sense, we can nevertheless think of them as being able to "set" their own values. For all players, then, the move 0 sets its value to 0 in the next world, while 1 sets its value to 1.

The above moves are sufficient for those variables that are exogenous, i.e., those whose values are not dependent on the values of any other variables. In our scenario, Q and R are exogenous variables. For an endogenous variable such as S, whose value is dependent on the values of Q and R, it is not possible to represent the causal law governing S, only using some combination of moves 0 and 1. The reason is because the value of S in the next world is dependent on the values of Q and R in the next world, not the current world. For endogenous variables, therefore, we introduce a third move x, which sets the value of the endogenous variable according to its structural equation. For example, the move x for player S sets the value of S in the next world to be equal to $\neg Q \wedge \neg R$. In summary, exogenous variables have two moves 0 and 1, while endogenous variables have a third move x.

The initial set of propositions is $P = \{q, r, s\}$. Our propositions correspond to valuations of each of the variables; e.g., q is true in those worlds where the value of Q is 1, etc. Where necessary, the values of endogenous variables can be calculated using their structural equations. For example, the value of S in w_{10x} is $\neg 1 \wedge \neg 0 = 0 \wedge 1 = 0$. The labeling function is then straightforward to calculate: $L(w_{000}) = \varnothing$, $L(w_{10x}) = \{q\}$, etc.

Finally, we must specify our initial conditions: the initial strategies of each player. For player S, the strategy is to enforce the causal law $S = \neg Q \wedge \neg R$ at each world. Therefore the strategy for S is simply $\lambda w.x$: at all worlds w, make move x.

As for players Q and R, because they are exogenous variables, they do not have structural equations in Pearl's causal models (Pearl, 2000). However, we do not want to say that they have no strategies. As previously mentioned, when evaluating a counterfactual sentence, we only want to consider those worlds that are closest to the actual world. But in ATL+I, having no strategy means placing no restrictions on which worlds are accessible from the actual world (Jamroga et al., 2005). Intuitively, given a world with some value of Q, worlds with the same value of Q 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 Q the same:

$$\sigma(Q) = \lambda w. \begin{cases} 1, & q \in L(w) \\ 0, & \text{otherwise} \end{cases}$$

The strategy for R can be similarly specified.

3 The dynamics of causal counterfactuals

Now we can describe the evaluation of counterfactual sentences in our framework. We translate sentences into formulas of type `Form`. In addition to the formulas of propositional and basic modal logic, we also include the formula scheme `Str a strategy phi`; these correspond to ATL+I sentences $(\mathbf{str}_a \sigma_a)\phi$. In ATL+I, it is the evaluation of $\mathbf{str}$-formulas in which changes of strategy occur; in our framework, counterfactual sentences are translated into $\mathbf{str}$-formulas for evaluation.

Formulas must of course be evaluated relative to some model. In addition, in a dynamic semantics, we must also keep track of the context. To do this, we make use of Haskell's state monad. We define the type `Model` of our states as a record type, that includes the current strategy function, as well as four components of our concurrent game structure: the sets of players and worlds, and the labeling and transition functions. Because of how we constructed our concurrent game structures above, the set of propositions and the move function can be inferred from the other components.

For a given function (and context), our model checker returns the set of possible worlds where the formula is true. As such, our main function, `check`, has type `Form -> State Model [World].` The model checker is based heavily on that in (Jamroga et al., 2005) for ATL+I, which itself is derived from the model checker for ATL in (Alur et al., 2002). Propositions are checked using the labeling function, and formulas of propositional logic follow via the usual set-theoretic operations. The checking of modal formulas makes use of a pre-image function, which, given a set of possible worlds, returns the set of worlds that can access any of those worlds. Then, for example, to check a formula $\Diamond\phi$, we first find the set of worlds where ϕ is true, and then calculate the set of worlds such that the ϕ-worlds are accessible.

Finally, to check **str**-formulas, we introduce a `revise` function. This is the mechanism by which causal interventions are modeled. Formally, let σ be the current strategy for player a, and σ' be a's new strategy. Then we can say that $\texttt{revise}(a, \sigma') = \{\sigma \cup \sigma'\}$.

We should note that our `revise` function differs from that of Jamroga et al. (2005). Whereas changes of strategy in ATL+I involve replacement of the player's previous strategy, our `revise` function simply add the moves from σ' to a's previous strategy. We recall that in von Fintel's dynamic account of counterfactuals, the accessibility relation (modal horizon) expands but does not contract. In other words, all worlds accessible from a given world before an update to the model, remain accessible afterwards.

Returning to the kangaroos, we can now see the difference in the evaluation of the Lewis-Sobel sequence in (2) and the reverse Sobel sequence in (3). We will use the propositions q, r, and s as before, to represent kangaroos having tails, using crutches, and toppling over, respectively. In evaluating the sentence "If kangaroos had no tails, they would topple over" under the causal modeling approach, we apply an intervention in the model to set $Q = 0$. This corresponds to a strategy for Q to go to a world where $\neg q$ is true; i.e., $\lambda w.0$. Then, following von Fintel (2001), we check whether in all accessible worlds where $\neg q$ is true, s is also true; this is the strict conditional $\Box(\neg q \rightarrow s)$. Therefore, the formula we want to evaluate is $(\mathbf{str}_Q(\lambda w.0))\Box(\neg q \rightarrow s)$.

Similarly, when we evaluate the sentence "If kangaroos had no tails but used crutches, they would not topple over", we want to expand our modal horizon to include worlds where $\neg q$ and r are both true. This involves changes in strategy by both Q and R; Q to set $Q = 0$, R to set $R = 1$. The formula to be evaluated must therefore include both an $(\mathbf{str}_Q(\lambda w.0))$ term and an $(\mathbf{str}_R(\lambda w.1))$ term. Then, since we want to check the truth of $\neg s$ in those accessible worlds where both $\neg q$ and r are true, our formula is $(\mathbf{str}_Q(\lambda w.0))(\mathbf{str}_R(\lambda w.1))\Box((\neg q \wedge r) \rightarrow \neg s)$.

Suppose that starting from our initial conditions, the sentence "If kangaroos had no tails, they would topple over" is uttered. We first update the strategy function for Q, to add the move 0 to Q's initial strategy. This has no effect in worlds where $Q = 0$, as the default strategy for Q is to keep its value the same. However, in worlds where $Q = 1$, Q now has two moves consistent with its new strategy, 0 and 1. Now, using the updated accessibility relation, we evaluate the formula $\Box(\neg q \rightarrow s)$. Every world now has an accessible $\neg q$-world. We note that according to the structural equation $S = \neg Q \wedge \neg R$, s will be true in those worlds where $\neg q$ and $\neg r$ hold. Since S's strategy is to enforce the structural equation, we know that it will hold in all accessible worlds. In addition, R's strategy continues to dictate that from every world, any accessible world will have the same valuation of R. We conclude that the sentence is true in those worlds where $R = 0$; these include the actual world w_{10x}.

Then suppose the sentence "If kangaroos had no tails but used crutches, they would not topple over" is uttered. Again, we update the strategy function for Q to add move 0. But since 0 was previously added when evaluating the first sentence, this revision has no effect. Next, we add the move 1 to all worlds in R's strategy, similarly as before. Now we evaluate the strict conditional $\Box((\neg q \wedge r) \rightarrow \neg s)$. Since S's strategy still has not changed, the causal law $S = \neg Q \wedge \neg R$ continues to hold in every accessible world. Therefore, for every world in our model, in every accessible world where $(\neg q \wedge r)$ is true, $\neg s$ is true, and so is the sentence.

What if the order of the two sentences were reversed? First, starting again from the initial conditions, the sentence "If kangaroos had no tails but used crutches, they would not topple over" is uttered. Because move 0 had not been added yet, it is this sentence that adds 0 to Q's strategy. All other effects of uttering this sentence are the same as before, as is the set of possible worlds where it is true. However, we can see a difference in the evaluation of the second sentence "If kangaroos had no tails, they would topple over". Updating the strategy function for Q has no effect, since the move 0 has already been added to Q's strategy by the first sentence. Furthermore, it is no longer the case that R's strategy keeps the valuation of R constant; as a result of the first sentence, move 1 is now available to R at every world. In other words, among the $\neg q$-worlds accessible from any given world, one of them will also be an r-world. Since $S = \neg Q \wedge \neg R$ holds in every world, we know that from any world, one of the accessible $\neg q$-worlds will not be an s-world. We conclude that the sentence does not hold in any world.

4 Discussion

4.1 Translating counterfactual sentences into str-formulas

One challenge in synthesizing a causal modeling approach to counterfactuals with a possible-worlds semantics is the difference in how counterfactual sentences are evaluated in the two approaches. Under the classic possible-worlds framework, we check whether in the closest possible worlds where the antecedent is true (making any changes to the accessibility relation, if necessary, to ensure that at least one such possible world exists), the consequent is true. In a causal theory of counterfactuals, the antecedent of the counterfactual determines the intervention to be applied to the causal model. Then, the consequent is evaluated relative to the new model.

In this paper, we identified the necessary change in the accessibility relation with the intervention in the causal model, which we implement as a change in strategy for some player. Such an approach raises two questions. The first question concerns which possible worlds count as worlds where the antecedent is true. In the kangaroo example, when we translated a counterfactual of the form $\phi > \psi$ into an $\mathtt{str}$-formula, the strict conditional portion of the formula was simply $\Box(\phi \rightarrow \psi)$. In other words, if the antecedent of the counterfactual is ϕ, then we check whether the accessible ϕ-worlds are also ψ-worlds. However, there is evidence that this approach may not work for all scenarios.

Briggs (2012) discusses the scenario, originally found in Pearl (2000), of an execution of a prisoner. A full description of the scenario can be found in either of the above papers; we note here that there are two executioners, X and Y, and whether they fire is determined by whether the captain C signals for them to do so. In other words, the behavior of executioner X is governed by the structural equation $X = C$. If either executioner fires, the prisoner dies. In the actual world, the captain signals, both executioners fire, and the prisoner dies.

Briggs considers the sentence "If executioner X had fired, then (even) if the captain had not signalled, the prisoner would have died." Under a causal model, we intervene to change the structural equation $X = C$ to $X = 1$. However, in the classic possible-worlds framework, no change in the accessibility relation is necessary. Executioner X fires in the actual world, and as a consequence of (weak) *centering*, the assumption that every world is at least as similar to itself as to any other world, every world is then accessible to itself. Under the classic approach, we check the truth of the consequent in the closest possible world where the antecedent is true; i.e., the actual world, where the consequent is false. But as Briggs notes, applying the intervention to the causal model changes the truth of the consequent.

When specifying a set of possible worlds corresponding to a causal model, we must distinguish between worlds where different causal laws hold. For example, in the kangaroo scenario, we distinguish worlds w_{10x} (where kangaroos do not topple over because they have tails) and w_{100} (where they do not topple over, because it is a law of nature that they never topple over), even though the same propositions are true in both worlds: $L(w_{100}) = L(w_{10x}) = \{q\}$. Likewise, the antecedent of the counterfactual in the execution case is the proposition that executioner X fires; let us call it x. We note that x does not determine what structural equation holds in a particular world; in some x-worlds, the relevant causal law is $X = 1$, while in others, it is $X = C$. When we say "if executioner X had fired" in a causal model, the relevant possible worlds are those in which the structural equation is $X = 1$. The corresponding proposition is not x, but a different proposition (call it x_1), which is true in exactly those worlds where the causal law $X = 1$ holds.

4.2 Counterfactuals with complex antecedents

Second, we note that antecedents of counterfactuals are propositions (of type `Prop`), while strategies have type `World -> [Move]`. Is there a way to systematically translate propositions into strategies? We have already seen that for atomic propositions such as r, we intervene to make sure that there is an accessible world where r is true, by adding the move 1 to the strategy of player R at every world: $\lambda w.1$. Similarly, for negations of atomic propositions, such as $\neg q$, we add move 0 to Q's strategy: $\lambda w.0$.

We have also seen an example of a conjunction, $(\neg q \wedge r)$. To ensure that there is an accessible world where the conjunction holds, we simply have both players change their strategies in sequence: $(\mathbf{str}_Q(\lambda w.0))(\mathbf{str}_R(\lambda w.1))$.... We note that the order each player changes their strategy does not matter. The moves each player is allowed to make are affected only by their own strategy, not those of any other players, and the strict conditional portion of the counterfactual formula is only evaluated after all strategy changes.

For other complex antecedents, Briggs (2012) borrows the idea of a *state space* from Fine (2012). States are defined by a valuation of some variable(s); e.g., $Q = 0 \wedge R = 1$. For propositional antecedents (including negations, conjunctions, disjunctions, and material conditionals), Briggs specifies states that make the antecedent true. For example, a disjunctive antecedent $(\phi \vee \psi)$ is made true by three states or interventions: one that sets $\phi = 1$, one that sets $\psi = 1$, and one that sets both $\phi = 1 \wedge \psi = 1$.

One challenge that arises in adapting this approach to ours is that evaluating the disjunction involves checking the results of three different interventions applied to the original model. However, in our dynamic semantics, once an intervention is made, the moves added to the player's strategy remain available to future evaluations; there is no "going back" to try a different intervention. In addition, while the states associated with the disjunction $(\phi \vee \psi)$ are the same as those associated with the negated conjunction $\neg(\neg\phi \wedge \neg\psi)$, Ciardelli et al. (2018) provide evidence that those antecedents in fact have different meanings.

Furthermore, it is not clear what impact, if any, a disjunctive antecedent should have on the accessibility relation at all. Ciardelli et al. (2018) discuss the example of two switches for a light. They are connected in such a way that the light is on if the switches are both up or both down, and off otherwise. In the actual world, the switches are both up and the light is on. While Ciardelli et al. do not consider sequences of counterfactuals, it is easy enough to construct a reverse Sobel sequence as with the kangaroos:

(4) If switch A and switch B were both down, the light would be on.
#If switch A was down, the light would be off.

Now let us replace the conjunction with a disjunction. In their experiment, Ciardelli et al. found that the sentence "If switch A or switch B was down, the light would be off." was judged by most participants to be true (in contrast with the sentence "If switch A and switch B were not both up, the light would be off.", with a negated conjunctive antecedent). If we use this sentence instead in our sequence, the infelicity seems to go away:

(5) If switch A or switch B was down, the light would be off.
If switch A was down, the light would be off.

In fact, according to the rule of simplification of disjunctive antecedents, the second sentence is a logical consequence of the first. Nevertheless, this indicates that perhaps the modal horizon did not expand to include worlds where switch B was down in this case, at least not permanently; if it had, then we would have to consider them when evaluating the second sentence. Alternatively, von Fintel (2001) suggests that logical arguments, unlike normal discourse, carry with them an assumption of constant context. Certainly more research is needed in this area.

5 Conclusion

In this paper, we present a semantics for counterfactuals that combines ideas from dynamic semantics and causal modeling approaches. Our implementation is based on concurrent game structures, where variables are interpreted as players and interventions as changes in players' strategies. Using the classic example of kangaroos with no tails, we show how our approach is able to capture judgments about sequences of counterfactuals.

Acknowledgements

We would like to thank the reviewers for their helpful comments.

References

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

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

Ciardelli, I., L. Zhang, and L. Champollion (2018, Dec). Two switches in the theory of counterfactuals. *Linguistics and Philosophy 41*(6), 577–621.

Fine, K. (2012). Counterfactuals without possible worlds. *The Journal of Philosophy 109*(3), 221–246.

Groenendijk, J. and M. Stokhof (1991). Dynamic predicate logic. *Linguistics and Philosophy 14*(1), 39–100.

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

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

Kripke, S. A. (1963). Semantical considerations on modal logic. *Acta Philosophica Fennica 16*(1963), 83–94.

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

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

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

Pollock, J. (1976). *Subjunctive reasoning*. Springer.

Stalnaker, R. C. (1968). A theory of conditionals. In W. L. Harper, R. Stalnaker, and G. Pearce (Eds.), *Ifs: Conditionals, Belief, Decision, Chance and Time*, pp. 41–55. Dordrecht: Springer Netherlands.

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