[5] T. Costello and J. McCarthy (1999) Useful counterfactuals. *Electronic Transactions on Artificial Intelligence* Volume 3.[3]

[6] H.A. Kautz (1986) The logic of persistence, Proc. AAAI'86, pp. 401-405.

[7] S.C. Kleene (1952) *Introduction to Metamathematics*. North-Holland, Amsterdam.

[8] D. Lewis (1973) *Counterfactuals*. Basil Blackwell, Oxford.

[9] D. Lewis (1979) Counterfactual dependence and time's arrow. *Noûs* 13, pp. 455-476.

[10] J.L. Mackie (1975) Causes and Conditions. *American Philosophical Quarterly* 2.4, 1965, pp. 245-255 and 261-264.

[11] J. McCarthy and P. Hayes (1969) Some philosophical problems from the standpoint of Artificial Intelligence. In *Machine Intelligence* 4, B. Meltzer and D. Michie (Eds.), Edinburgh University Press, Edinburgh.

[12] J. Pearl (1996) Causation, action, and counterfactuals. Proc. TARK VI, pp. 51-73.

[13] J. Pearl (1999) Reasoning with cause and effect. Proc. IJCAI-99, pp. 1437-1449.

[14] Y. Shoham (1988) *Reasoning About Change*, M.I.T. Press, Cambridge Mass.

[15] R.C. Stalnaker (1975) A theory of conditionals. In *Causation and Conditionals*, E. Sosa (ed.), Oxford University Press, Oxford, 1975, pp. 165-179. First published in N. Rescher (ed.) *Studies in Logical Theory*, Basil Blackwell, Oxford, 1968.

[16] R.C. Stalnaker and R.H. Thomason (1970) A semantic analysis of conditional logic. *Theoria* 36, pp. 23-42.

[17] F. Veltman (1985) *Logics for Conditionals*. Ph.D. thesis, University of Amsterdam.

[18] G. White, J. Bell and W. Hodges (1998) Building Models of Prediction Theories. Proc. KR'98, pp. 557-568.

[3] Available at: www.ep.liu.se/ea/cis/1999/012/.

is false.

Proof. At w_2 in the previous proof, $\neg$*Succ*$(Kill(O,K))(1)$ is true and $\exists x($*Succ*$(Kill(x,K))(1) \wedge x \neq O)$ is false. $\Diamond$

Alternatively, suppose that Oswald succeeded in killing Kennedy at time 1, and that someone else attempted to kill Kennedy at time 1:

$$\Theta_{11.3} = \Theta_{11} \cup \{\textit{Succ}(Kill(O,K))(1), \\ \exists x(\textit{Occ}(Kill(x,K))(1) \wedge x \neq O)\}.$$

Then $\Theta_{11.3}$ predicts:

$$\neg\textit{Succ}(Kill(O,K))(1) \Downarrow \\ \exists x(\textit{Succ}(Kill(x,K))(1) \wedge x \neq O)$$

Proof. $\Theta_{11.3}$ is non-deterministic. Any actual world w_0 in a causal model for $\Theta_{11.3}$ can be represented as follows:

$$\{Alive(K)(1), \textit{Occ}(Kill(O,K))(1), \\ \textit{Occ}(Kill(A,K))(1), \neg Alive(K)(2), \ldots\}.$$

where A denotes some object in the domain other than that denoted by O. Note that since *Succ*$(Kill(O,K))(1)$ is true at w_0, it follows from the law of change and inertia and the definition of *NCause* that *Qual*$(Kill(A,K))(1)$ is true at w_0; Oswald's success preempts that of would-be assassin A.

Let w_1 be a closest world below w_0 at which $\bullet$*Succ*$(Kill(O,K))(1)$, is true. Then w_1 can be represented as follows:

$$\{Alive(K)(1), \textit{Occ}(Kill(A,K))(1), \neg Alive(K)(2), \ldots\}.$$

As in the penultimate proof, *Occ*$(Kill(O,K))(1)$ is undefined at w_1, and consequently assassin A now succeeds.

Let w_2 be a closest world above w_1 at which $\neg$*Succ*$(Kill(O,K))(1)$ is true. Then w_2 can be represented as:

$$\{Alive(K)(1), \neg\textit{Occ}(Kill(O,K))(1), \\ \textit{Occ}(Kill(A,K))(1), \neg Alive(K)(2), \ldots\}.$$

or as:

$$\{Alive(K)(1), \textit{Occ}(Kill(O,K))(1), \textit{Qual}(Kill(O,K))(1) \\ \textit{Occ}(Kill(A,K))(1), \neg Alive(K)(2), \ldots\}.$$

In either case $\exists x($*Occ*$(Kill(x,K))(1) \wedge x \neq O)$ is true at w_2, and consequently so is $\exists x($*Succ*$(Kill(x,K))(1) \wedge x \neq O)$.

It follows that:

$$\neg\textit{Succ}(Kill(O,K))(1) \Uparrow \\ \exists x(\textit{Succ}(Kill(x,K))(1) \wedge x \neq O)$$

is true at w_1. So the required contrafactual is true at w_0. $\Diamond$

Clearly also $\Theta_{11.3}$ predicts that the contrafactual:

$$\neg\textit{Succ}(Kill(O,K))(1) \Downarrow \\ \neg\exists x(\textit{Succ}(Kill(x,K))(1) \wedge x \neq O)$$

is false.

Proof. At w_2 in the previous proof, $\neg$*Succ*$(Kill(O,K))(1)$ is true and $\neg\exists x($*Occ*$(Kill(x,K))(1) \wedge x \neq O)$ is false. $\Diamond$ $\Box$

7 Concluding remarks

This paper has proposed a formal theory of causal counterfactuals which combines a partial-worlds semantics for counterfactuals with a formal pragmatics for common sense reasoning about actual events, in order to provide a unified framework for formal reasoning about actual and counterfactual events.

The question of the logic of the new conditionals is an interesting one. However it seems that finding an axiomatization (and an appropriate proof-theoretic pragmatics for the axioms of causal theories) is not a necessary preliminary to implementation. The causal models for a causal theory Θ differ only in inessential detail. In practice, it is possible to fix a single, canonical, interpretation of those components of models which do not figure in the definition of the equivalence relation $\sim_\Theta$; thus time is taken to be isomorphic to the natural numbers (or the integers, etc.) and temporal terms are interpreted accordingly, the objects denoted by the names of material objects are fixed ("symbol grounding"), etc. The canonical interpretation results in the $\sim_\Theta$-relation containing a single equivalence class. Moreover, this class contains a $\sim_\Theta$-maximal element, which consists of the "union" of all of the models in the class, and which can be called the canonical causal model for Θ. Consequently it seems that the model-building approach described in [18] can be extended, and that appropriate parts of the canonical causal model for a causal theory can be constructed chronologically. The evolving partial models of causal theories in [18] provide the actual world(s) of the model, and it seems that these can be used as a basis for building sufficient counterfactual worlds in order to evaluate the new conditionals correctly.

It will also be interesting to apply the present theory to the formalization of reasoning about plans as suggested in the introduction; this may involve adding temporal intervals and complex events to the language of $\mathcal{MTC}$, but doing so is straightforward.

Another development will be a theory of *intentional conditionals*. This will combine the semantics for conditionals given here with an appropriate formal theory for reasoning about actual intentional states (such as beliefs, goals and obligations), in order to provide a formal analysis of counterfactuals such as "If John had known that Mary had done α, then he would have done β".

References

[1] E. Adams (1970) Subjunctive and indicative conditionals. *Foundations of Language* 6, pp. 89-94.

[2] J. Bell (2000) Primary and Secondary Events. Submitted to *Electronic Transactions on Artificial Intelligence*.[2]

[3] J. Bennett (1974) Review of [8]. *The Canadian Journal of Philosophy* 4, pp. 381-402.

[4] J.P. Burgess (1981) Quick completeness proofs for some logics of conditionals. *Notre Dame Journal of Formal Logic* 22(1), 1981, pp. 76-84.

[2] Available at: www.ida.liu.se/ext/etai/received/actions/sframe.html.

did. Accordingly, an indicative analysis of the conditional can be given. Let $\Theta_{11.1} = \Theta_{11} \cup \{\neg Alive(K)(2)\}$. Then $\Theta_{11.1}$ predicts:

$$\neg \textit{Succ}(Kill(O, K))(1) \rightarrow \\ \exists x(\textit{Succ}(Kill(x, K))(1) \wedge x \neq O)$$

Proof. Let w_0 be a chronologically minimial $\Theta_{11.1}$-world in a causal model for $\Theta_{11.1}$. By the inertia axiom and the law of change and inertia, $\exists e \textit{Cause}(\textit{Occ}(e)(1), \bullet Alive(K)(2))$ is true at w_0. So, by the definition of *SCause*, there must be some individual A such that $\textit{Succ}(Kill(A, K))(1)$ is true at w_0. If $\neg \textit{Succ}(Kill(O, K))(1)$ is true at w_0, then it must also be the case that $A \neq O$. Consequently the desired conditional is true at w_0.

Note that $\Theta_{11.1}$ is non-deterministic; there may be many candidates for the actual world in a causal model for $\Theta_{11.1}$. This is appropriate: if the assassin A is not Oswald, then A's identity is unknown. $\Diamond$

By contrast, the second conditional is counterfactual. It is evaluated on the basis that Oswald did kill Kennedy. Thus if Oswald acted alone it should be false, while if there was another assassin it should be true. Similarly the third conditional is counterfactual. Given that Oswald did kill Kennedy, it should be true if he acted alone, and false if there was another assassin.

These counterfactuals were used by Lewis [8, p. 71] against metalinguistic theories of counterfactuals, and in turn by Bennett [3] to raise the "future similarity" objection to Lewis' own analysis [8]; recall from Section 2 that this depends on the notion of comparative overall similarity, with (classical) possible worlds being ordered according to their comparative overall similarity to the actual world.

Lewis believes that Oswald killed Kennedy and that he acted alone. Consequently he considers that the first of the pair of Oswald-Kennedy counterfactuals is false and that the second is true; [8, p. 3, p. 71]. However, Bennett objects that among the worlds in which Oswald did not kill Kennedy, the worlds in which someone else did seem to be more similar to the actual world than those worlds in which noone else did; a world in which a Dallas policeman decided to kill Kennedy on the spur of the moment and in which the course of events then reconverged with that of the actual world seems to be more similar to the actual world than a world in which Kennedy was not killed and the course of events continued to diverge from that of the actual world thereafter. So, according to Lewis' analysis and contrary to his opinion, it seems that the first of the Oswald-Kennedy counterfactuals should be true and that the second should be false.

Lewis counters that we need to respect the "extreme shiftiness and context-dependence of similarity" and be careful to distinguish between "the similarity relations that guide our offhand explicit [similarity] judgements and those that govern our counterfactuals in various contexts" [9, p. 466]. He then proceeds to develop a set of constraints which further restrict the choice of comparative similarity relations for counterfactuals such as these. However these are not formal and have proved contentious.

By contrast, the formal pragmatics of causal theories has been developed in response to the extreme shiftiness and context-dependence of common sense causal reasoning. Its use as a formal pragmatics for causal counterfactuals results in the correct evaluation of the Oswald-Kennedy counterfactuals, and, more generally, it is evident that it is not prone to the future-similarity objection.

In order to see this, suppose firstly that Oswald succeeded in killing Kennedy at time 1, and that noone else attempted to kill Kennedy at time 1:

$$\Theta_{11.2} = \Theta_{11} \cup \{\textit{Succ}(Kill(O, K))(1), \\ \neg \exists x(\textit{Occ}(Kill(x, K))(1) \wedge x \neq O)\}.$$

Then $\Theta_{11.2}$ predicts:

$$\neg \textit{Succ}(Kill(O, K))(1) \Downarrow \\ \neg \exists x(\textit{Succ}(Kill(x, K))(1) \wedge x \neq O)$$

Proof. $\Theta_{11.2}$ is deterministic. The actual world, w_0, in a causal model for $\Theta_{11.2}$ can be represented as follows:

$$\{Alive(K)(1), \textit{Occ}(Kill(O, K))(1), \\ \neg \textit{Occ}(Kill(A_i, K))(1), \neg Alive(K)(2), \ldots\}.$$

where $\neg \textit{Occ}(Kill(A_i, K))(1)$ represents every literal of the form $\neg \textit{Occ}(Kill(A, K))(1)$ which is such that A and O denote different domain objects.

Let w_1 be the closest world below w_0 at which $\circ \neg \textit{Succ}(Kill(O, K))(1)$, or equivalently $\bullet \textit{Succ}(Kill(O, K))(1)$, is true. Then w_1 can be represented as follows:

$$\{Alive(K)(1), \neg \textit{Occ}(Kill(A_i, K))(1), Alive(K)(2), \ldots\}.$$

In order to see that the atom $\textit{Occ}(Kill(O, K))(1)$ is undefined at w_1, suppose that this is not the case. Then, as $w_1 \prec_P w_0$, it follows by Proposition 9, that the atom is true at w_1. But then, as the precondition $Alive(K)(1)$ is true at w_1, it follows from the truth of $\bullet \textit{Succ}(Kill(O, K))(1)$ that $\textit{Qual}(Kill(O, K))(1)$ is true at w_1. But then by Proposition 9 it follows that $\textit{Qual}(Kill(O, K))(1)$ is true at w_0, and this contradicts the truth of $\textit{Succ}(Kill(O, K))(1)$ at w_0.

Let w_2 be a closest world above w_1 at which $\neg \textit{Succ}(Kill(O, K))(1)$ is true. Then w_2 can be represented as:

$$\{Alive(K)(1), \neg \textit{Occ}(Kill(O, K))(1), \\ \neg \textit{Occ}(Kill(A_i, K))(1), Alive(K)(2), \ldots\}.$$

or as:

$$\{Alive(K)(1), \textit{Occ}(Kill(O, K))(1), \textit{Qual}(Kill(O, K))(1), \\ \neg \textit{Occ}(Kill(A_i, K))(1), Alive(K)(2), \ldots\}.$$

In either case, $\neg \exists x(\textit{Occ}(Kill(x, K))(1) \wedge x \neq O)$ is true at w_2, and consequently so is $\neg \exists x(\textit{Succ}(Kill(x, K))(1) \wedge x \neq O)$.

It follows that:

$$\neg \textit{Succ}(Kill(O, K))(1) \Uparrow \\ \neg \exists x(\textit{Succ}(Kill(x, K))(1) \wedge x \neq O)$$

is true at w_1. So the required contrafactual is true at w_0. $\Diamond$

Clearly also $\Theta_{11.2}$ predicts that the contrafactual:

$$\neg \textit{Succ}(Kill(O, K))(1) \Downarrow \\ \exists x(\textit{Succ}(Kill(x, K))(1) \wedge x \neq O)$$

the lot. As this is all that is known, it is reasonable to predict that the car will still be there at time 2. Given that it is, and again that nothing else is known, it is again reasonable to predict that the car will still be there at time 3.

Nevertheless, given that the car has gone at time 4, it also seems reasonable to maintain that the car *could* have been stolen at time 1 or time 2 or time 3; as any of these events, together with inertia, would explain the car's absence at time 4.

This apparent paradox can be resolved by distinguishing between predicting on the basis of the current state of the actual world, and retrospectively seeking all of the reasonable explanations for some given aspect of the actual world. Given that the car is in the lot at time 1 and that is is gone at time 4, the explanatory problem is to produce all of the reasonable explanations for the car's disappearance. It is thus necessary to consider all of the events which *might* have caused it. The emphasized modalities in the above discussion suggest that explanatory reasoning is a form of counterfactual reasoning. Moreover, as the events in question should be compatible with what is known of the actual world, it suggests that complefactuals can be used to generate explanations. And, indeed, assuming that a finite number of events are involved, it is possible to use complefactuals in order to give a formal definition of explanation.

The occurrence of event e at time t is *an* explanation for ϕ at the later time t' at a world w if and only if the closest worlds above w in which e occurs at t: e causes ϕ at $t+1$, and no subsequent event e' occurring before time t' causes $\bullet\phi$:[1]

$$\begin{aligned} &\mathit{Expl}(\mathit{Occ}(e)(t), \phi(t')) \equiv \\ &\quad t < t' \wedge \\ &\quad (\mathit{Occ}(e)(t) \Uparrow (\mathit{Cause}(\mathit{Occ}(e)(t), \phi(t+1)) \wedge \\ &\qquad \bullet \exists e', t''(t < t'' < t' \wedge \\ &\qquad\qquad \mathit{Cause}(\mathit{Occ}(e')(t''), \bullet\phi(t''+1)))) \end{aligned} \tag{12}$$

Note that a weaker notion of explanation based on sufficient causation can be provided using *SCause* instead of *Cause*.

As there are a finite number of such explanations, they can be collected together as a disjunction representing *the* explanation:

$$\begin{aligned} &\mathit{Expls}(\phi, \psi(t)) \equiv \\ &\quad (\phi \equiv \bigvee \{\mathit{Occ}(e)(t') : \mathit{Expl}(\mathit{Occ}(e)(t'), \psi(t))\}) \end{aligned} \tag{13}$$

These two definitions form a theory of explanation, Θ_E.

Let $\Theta_{10.2} = \Theta_E \cup \Theta_{10.1}$, and let M be a causal model for $\Theta_{10.2}$. Then, as required, $\Theta_{10.2} \mid\!\approx^M_P \mathit{Expls}(\mathit{Occ}(Steal)(1) \vee \mathit{Occ}(Steal)(2) \vee \mathit{Occ}(Steal)(3), \neg In(4))$.

Proof. The closest world above w_0 at which $\mathit{Occ}(Steal)(3)$ is true can be represented as follows:

$$\{In(1), In(2), In(3), \mathit{Occ}(Steal)(3), \neg In(4), \ldots\}$$

Indeed, this world is w_0 itself. As in the previous proof, $In(3)$ is true at w_0. And, as w_0 is a $\Theta_{10.2}$-world, $\neg In(4)$ is true at w_0. So it follows from the contrapositive of the inertia axiom that $\mathit{Aff}(In)(3)$ is true at w_0. As this is the case, it follows from the law of change and inertia that $\exists e \mathit{Cause}(\mathit{Occ}(e)(3), \bullet In(4))$ is true at w_0. Moreover, it follows from the definition of *SCause* that $e = Steal$. So $\mathit{Cause}(\mathit{Occ}(Steal)(3), \neg In(4))$ is true at w_0. Clearly also $\mathit{Expl}(\mathit{Occ}(Steal)(3), \neg In(4))$ is true at w_0.

The closest world above w_0 at which $\mathit{Occ}(Steal)(2)$ is true, call it w_1, can be represented as follows:

$$\{In(1), In(2), \mathit{Occ}(Steal)(2), \neg In(3), \neg In(4), \ldots\}.$$

At any time point qualifications are minimized before affectations (change is preferred to inertia). Thus the steal event succeeds at time 2 and its postconditions are true at time 3. The definition of $\prec_P$ ensures that no other events occur at time 2 or subsequently. So $\mathit{Cause}(\mathit{Occ}(Steal)(2), \neg In(3))$ is true at w_1 and $\mathit{Expl}(\mathit{Occ}(Steal)(2), \neg In(4))$ is true at w_0.

Similarly, the closest world above w_0 at which $\mathit{Occ}(Steal)(1)$ is true, call it w_2, can be represented as follows:

$$\{In(1), \mathit{Occ}(Steal)(1), \neg In(2), \neg In(3), \neg In(4), \ldots\}.$$

The definition of $\prec_P$ ensures that no other events occur at time 1 or subsequently. So $\mathit{Cause}(\mathit{Occ}(Steal)(1), \neg In(2))$ is true at w_2, and $\mathit{Expl}(\mathit{Occ}(Steal)(1), \neg In(4))$ is true at w_0.

Note that $w_0 \prec_P w_1 \prec_P w_2$ and that the steal events occur at progressively earlier times at the worlds along this chain. Note also that $\Theta_{10.2}$ is deterministic. Any world in M which differs from w_0 is not a chronologically minimal world for $\Theta_{10.2}$. $\Diamond\,\Box$

In this example alternative worlds could be found which diverged from the actual world at some earlier point in its history, and consequently complefactuals could be used to refer to such worlds. The next example illustrates the use of contrafactuals when referring to worlds which have a common history with the actual world, but which differ from it at the present moment.

Example 11. The following conditionals have been much discussed:

> If Oswald did not kill Kennedy, then someone else did.
>
> If Oswald had not killed Kennedy, then someone else would have.
>
> If Oswald had not killed Kennedy, then noone else would have.

They will be discussed here in the context of the causal theory $\Theta_{11} = \Theta_C \cup \{(14), (15), (16)\}$; where:

$$\mathit{Alive}(K)(1) \tag{14}$$

$$\mathit{Pre}(\mathit{Kill}(x,y))(t) \equiv \mathit{Alive}(y)(t) \tag{15}$$

$$\mathit{Post}(\mathit{Kill}(x,y))(t) \equiv \neg\mathit{Alive}(y)(t) \tag{16}$$

The first two conditionals are given by Adams [1] in order to illustrate the distinction between indicative conditionals and subjunctive (counterfactual) conditionals.

Given that the actual world is one in which Kennedy was killed, the first conditional is indicative: someone killed Kennedy, so if Oswald did not kill him, then someone else

[1] The notation $\phi(t)$ is used to denote any formula in which the temporal variable t occurs free.

to the world. Having thus fixed the facts and events at t, assumptions are made which affect the world's future. First it is assumed that the events occurring at t succeed if there is no evidence to the contrary. Then it is assumed that the facts which hold at t persist if there is no evidence to the contrary. Clearly it is reasonable to fix the facts and events at t before speculating about the future, as such speculations should not alter the present. It also seems natural to assume that events succeed before assuming that facts persist, to prefer change to inertia; see the examples in Section 6. The final clause in the definition of $\prec_P$ simply restricts the higher-order atoms which hold at the world at t to those which follow from the interpretation of Θ.

If there is a single chronologically minimal world for Θ in each causal model for Θ, then Θ can be considered to be *deterministic*; as Θ uniquely determines which world is the actual world. Otherwise Θ is non-deterministic, and any one of the chronologically minimal Θ-worlds in the model may be the actual world.

Now, the pragmatic consequences of Θ in a causal model for Θ can be defined to be the consequences of Θ in all chronologically minimal Θ-worlds in the model, and the pragmatic consequences of Θ simpliciter can be defined to be those which follow in all such worlds in all of the causal models for Θ.

Definition 8. *Let Θ be a causal theory with causal model M, and let ϕ be a sentence. Then* Θ predicts ϕ *relative to M, written $\Theta \mid\!\approx_P^M \phi$, if ϕ is true in all chronologically minimal worlds for Θ in M, and* Θ predicts ϕ, *written $\Theta \mid\!\approx_P \phi$, if $\Theta \mid\!\approx_P^{M'} \phi$ for every causal model M' of Θ.*

A example of the use of this pragmatics for reasoning about actual events (events which occur in the actual world) is given in the next section; many further examples from [2] are readily adapted.

6 Counterfactual events

In addition to providing a pragmatics for reasoning about actual events, causal models also provide a pragmatics for reasoning about counterfactual events. The definition of $\prec_P$ has the effect that as one proceeds up any chain of worlds above a chronologically minimal Θ-world w in a causal model for Θ one encounters worlds in which additional literals are true. Moreover these additions arise antichronologically; the higher one ascends, the earlier the additions. Thus the worlds above w represent the various ways in which the history of w would differ, in accordance with the laws of Θ, as a result of additions introduced at progressively earlier moments. Similarly, as one proceeds down a chain of worlds below w one encounters worlds in which fewer literals are true, and in which the deletions are antichronlogical. Thus the worlds below w represent the various ways in which the history of w would differ, in accordance with the laws of Θ, as a result of deletions made at progressively earlier moments.

Thus in addition to the "horizontal" persistence by default of Kleene literals from one time point to the next within a world, there is also a limited form of the "vertical" persistence of the information models of Section 3. Let a temporal literal be any first-order literal of the form $\sigma r(u_1, \ldots, u_n)(t)$ or higher-order literal of the form $\sigma hr(e_1, \ldots, e_n, \ell_1, \ldots, \ell_m)(t)$; where σ is either the strong negation operator ($\neg$) or the null string. Moreover, for model M, world w in M and time point t, let $M, w/t$ be the set of temporal literals which are true at w up to t; that is:

$$M, w/t = \{\Lambda(t') : t' \leq t \text{ and } M, w \models \Lambda(t')\}$$

Then the following limited form of vertical persistence obtains in causal models:

Proposition 9. *Let M be a causal model, and let w and w' be worlds in M such that $w \prec_P w'$. Then there is a time point t such that:*

$$M, w/t-1 = M, w'/t-1 \text{ and } M, w/t \subset M, w'/t.$$

Thus, if in a causal model M, $w \prec_P w'$ is true, then the worlds w and w' have a common history up to some time point t, and every temporal literal $\Lambda(t)$ which is true at w is also true at w'. The $\prec_P$-closest worlds above a world w at which an additional temporal literal $\Lambda(t)$ is true can thus be regarded as the most similar worlds to w at which $\Lambda(t)$ is true, as these worlds share a common history with w up to time t and otherwise differ minimally from w in that they are governed by the laws of Θ thereafter. Similarly the $\prec_P$-closest worlds below w in which a temporal literal $\Lambda(t)$ is no longer true can be regarded as the most similar worlds to w in which $\Lambda(t)$ is not true.

Two examples of the use of counterfactuals are now given. The first illustrates actual predictive reasoning and the use of the complefactual conditional in order to provide alternative possible explanations.

Example 10. The stolen car problem, suggested by Kautz [6], is commonly believed to be a decisive objection to the principle of chronological minimization. A car is in a parking lot at time 1, but is no longer there when its owner returns at time 4. As chronological minimization has the effect of delaying change, it will result in the prediction that the car is still in the lot at time 3. But this seems to be contrary to intuition, as the car could have been stolen at time 1 or time 2.

This example can be represented by the theory $\Theta_{10.1} = \Theta_C \cup \{(9), (10), (11)\}$; where:

$$In(1) \wedge \neg In(4) \tag{9}$$

$$\mathit{Pre}(Steal)(t) \equiv In(t) \tag{10}$$

$$\mathit{Post}(Steal)(t) \equiv \neg In(t) \tag{11}$$

Then, indeed, $\Theta_{10.1}$ predicts that the car will still be in the lot at time 3: for any causal model M for $\Theta_{10.1}$, $\Theta_{10.1} \mid\!\approx_P^M In(3)$.

Proof. $\Theta_{10.1}$ Let w_0 be a chronologically minimal $\Theta_{10.1}$-world in a causal model M for $\Theta_{10.1}$. Then $In(1)$ is true at w_0, as w_0 is a $\Theta_{10.1}$-world. It is consistent to assume that $?\mathit{Aff}(In)(1)$ is true at w_0, so it follows by the inertia axiom that $In(2)$ is true at w_0. Similarly, $In(3)$ is true at w_0. $\Diamond$

It is difficult to disagree with these predictions if one reasons in the evolving partial epistemic context of the actual world. At time 1 all that is known is that the car is parked in

$$
\begin{array}{lll}
M,w,g \models \psi \Uparrow \chi & \text{iff} & M,w,g \models \circ\psi \text{ and for every } w' \text{ such that } w \preceq w' \text{ and } M,w',g \models \psi \\
 & & \text{there is a } w'' \text{ such that } w \preceq w'' \preceq w' \text{ and } M,w'',g \models \psi \\
 & & \text{and for every } w''' \text{ such that } w \preceq w''' \preceq w'', M,w''',g \models \psi \rightarrow \chi \\
M,w,g \models\!\!\!\models \psi \Uparrow \chi & \text{iff} & M,w,g \models \circ\psi \text{ and there is a } w' \text{ such that } w \preceq w' \text{ and } M,w',g \models \psi \wedge \neg\chi \\
 & & \text{and there is no } w'' \text{ such that } w \preceq w'' \preceq w' \text{ and } M,w'',g \models \psi \wedge \chi \\
M,w,g \models \psi \Downarrow \chi & \text{iff} & M,w,g \models \neg\psi \text{ and for every } w' \text{ such that } w' \preceq w \text{ and } M,w',g \models \circ\psi \\
 & & \text{there is a } w'' \text{ such that } w' \preceq w'' \preceq w \text{ and } M,w'',g \models \circ\psi \\
 & & \text{and for every } w''' \text{ such that } w'' \preceq w''' \preceq w, M,w''',g \models \circ\psi \rightarrow (\psi \Uparrow \chi) \\
M,w,g \models\!\!\!\models \psi \Downarrow \chi & \text{iff} & M,w,g \models \neg\psi \text{ and there is a } w' \text{ such that } w' \preceq w \text{ and } M,w',g \models \circ\psi \wedge \neg(\psi \Uparrow \chi) \\
 & & \text{and there is no } w'' \text{ such that } w' \preceq w'' \preceq w \text{ and } M,w'',g \models \circ\psi \wedge (\psi \Uparrow \chi) \\
M,w,g \models \psi \Rightarrow \chi & \text{iff} & M,w,g \models \psi \Uparrow \chi \text{ or } M,w,g \models \psi \Downarrow \chi \\
M,w,g \models\!\!\!\models \psi \Rightarrow \chi & \text{iff} & M,w,g \models\!\!\!\models \psi \Uparrow \chi \text{ or } M,w,g \models\!\!\!\models \psi \Downarrow \chi
\end{array}
$$

Table 2: Satisfaction and violation conditions for complefactuals, contrafactuals and counterfactuals (see Definition 2).

$$
\begin{array}{lll}
M,w,g \models \psi \Uparrow \chi & \text{iff} & M,w,g \models \circ\psi \text{ and } \uparrow(\psi,w,g) \subseteq [\![\chi]\!]^M_g \\
M,w,g \models\!\!\!\models \psi \Uparrow \chi & \text{iff} & M,w,g \models \circ\psi \text{ and } \uparrow(\psi,w,g) \bullet [\![\neg\chi]\!]^M_g \\
M,w,g \models \psi \Downarrow \chi & \text{iff} & M,w,g \models \neg\psi \text{ and } \downarrow(\circ\psi,w,g) \subseteq [\![\psi \Uparrow \chi]\!]^M_g \\
M,w,g \models\!\!\!\models \psi \Downarrow \chi & \text{iff} & M,w,g \models \neg\psi \text{ and } \downarrow(\circ\psi,w,g) \bullet [\![\neg(\psi \Uparrow \chi)]\!]^M_g
\end{array}
$$

Table 3: Simplified satisfaction and violation conditions for complefactuals and contrafactuals (see Definition 4).

In order to define the causal models for causal theories, it is sufficient to specify appropriate world frames for them.

The worlds in a causal model for a causal theory Θ should satisfy the laws of Θ. Thus, if $\mathit{Laws}(\Theta) = \Theta_C \cup \{\mathit{Pre}(e)(t) \equiv \phi \in \Theta\} \cup \{\mathit{Post}(e)(t) \equiv \phi \in \Theta\}$, then a causal model for Θ should be a $\mathit{Laws}(\Theta)$-model; that is, all of the worlds in the model should be $\mathit{Laws}(\Theta)$-worlds. In order to ensure that all such worlds are considered, a causal model should be one in which the world set is otherwise maximal. These requirements are formalized as follows. The models M and M' are Θ-equivalent, written $M \sim_\Theta M'$, if and only if M and M' are $\mathit{Laws}(\Theta)$-models, and M and M' differ at most on world frames or the interpretation of relations; that is, M and M' agree except perhaps on their respective components $\langle \mathcal{W}, \prec, \mathcal{R}, \mathcal{HR}, \mathcal{V}_\mathcal{R}, \mathcal{V}_\mathcal{HR} \rangle$ and $\langle \mathcal{W}', \prec', \mathcal{R}', \mathcal{HR}', \mathcal{V}'_\mathcal{R}, \mathcal{V}'_\mathcal{HR} \rangle$. Then a model M with world set $\mathcal{W}$ is said to be $\sim_\Theta$ maximal if and only if for every model M' with world set $\mathcal{W}'$ such that $M \sim_\Theta M'$, it is the case that $\mathcal{W}' \subseteq \mathcal{W}$.

The closeness relation on a maximal world set $\mathcal{W}$ is based on the principle of prioritized chronological minimization; a refinement of the form of chronological minimization suggested by Shoham [14] which is discussed further in [2]. The relation $\prec_P$ partially orders worlds on the basis of (a particular form of) information growth over time. Thus if $w \prec_P w'$ then w is, in the appropriate sense, chronologically less defined than w'. Let $w \prec_P w'$ if and only if w and w' are worlds in $\mathcal{W}$ which agree on the interpretation of all relations up to some time point t, and:

- at least one more atom of the form $r(u_1, \ldots, u_n)(t)$ or $\mathit{Occ}(e)(t)$ is defined (is either true or false) at w', or
- w and w' agree on all of the above atoms, and at least one more atom of the form $\mathit{Qual}(e)(t)$ is defined at w', or
- w and w' agree on all of the above atoms, and at least one more atom of the form $\mathit{Aff}(\ell)(t)$ is defined at w', or
- w and w' agree on all of the above atoms, and at least one more atom of the form $hr(e_1, \ldots, e_n, \ell_1, \ldots, \ell_m)(t)$ is defined at w'.

Definition 6. *An $\mathcal{MTC}$-model M with world-frame $\langle \mathcal{W}, \prec \rangle$ is said to be* a causal model *for a causal theory Θ if M is $\sim_\Theta$ maximal and $\prec$ is the order $\prec_P$ on $\mathcal{W}$.*

The selected worlds in a model for a causal theory Θ are the chronologically least defined Θ-worlds.

Definition 7. *Let Θ be a causal theory with causal model M. Then a world w in M is a* chronologically minimal world for Θ *in M if $M, w \models \Theta$ and there is no other world w' in M such that $M, w' \models \Theta$ and $w' \prec_P w$.*

The chronologically minimial worlds for Θ are chosen because each represents an interpretation of Θ which can be regarded as being constructed chronologically and parsimoniously. At each time point t only those facts and events which follow from the earlier interpretation of Θ are added

$$
\begin{array}{rll}
M,w,g \models t < t' & \text{iff} & \mathcal{V}_g(t) \prec_T \mathcal{V}_g(t') \\
M,w,g \not\models t < t' & \text{iff} & \mathcal{V}_g(t) \not\prec_T \mathcal{V}_g(t') \\
M,w,g \models u = u' & \text{iff} & \mathcal{V}_g(u) \text{ is } \mathcal{V}_g(u') \\
M,w,g \not\models u = u' & \text{iff} & \mathcal{V}_g(u) \text{ is not } \mathcal{V}_g(u') \\
M,w,g \models r(u_1,\ldots,u_n)(t) & \text{iff} & \mathcal{V}_{\mathcal{R}}(r,w,\mathcal{V}_g(t))(\mathcal{V}_g(u_1),\ldots,\mathcal{V}_g(u_n)) = \textit{true} \\
M,w,g \not\models r(u_1,\ldots,u_n)(t) & \text{iff} & \mathcal{V}_{\mathcal{R}}(r,w,\mathcal{V}_g(t))(\mathcal{V}_g(u_1),\ldots,\mathcal{V}_g(u_n)) = \textit{false} \\
M,w,g \models v(t) & \text{iff} & v \in \mathcal{V}_L \text{ and } M,w,g \models \mathcal{V}_g(v)(t) \\
M,w,g \not\models v(t) & \text{iff} & v \in \mathcal{V}_L \text{ and } M,w,g \not\models \mathcal{V}_g(v)(t) \\
M,w,g \models hr(e_1,\ldots,e_n,\ell_1,\ldots,\ell_m)(t) & \text{iff} & \mathcal{V}_{\mathcal{HR}}(hr,w,\mathcal{V}_g(t))(\mathcal{V}_g(e_1),\ldots,\mathcal{V}_g(e_n),\mathcal{V}_g(\ell_1),\ldots,\mathcal{V}_g(\ell_m)) = \textit{true} \\
M,w,g \not\models hr(e_1,\ldots,e_n,\ell_1,\ldots,\ell_m)(t) & \text{iff} & \mathcal{V}_{\mathcal{HR}}(hr,w,\mathcal{V}_g(t))(\mathcal{V}_g(e_1),\ldots,\mathcal{V}_g(e_n),\mathcal{V}_g(\ell_1),\ldots,\mathcal{V}_g(\ell_m)) = \textit{false} \\
M,w,g \models \neg\psi & \text{iff} & M,w,g \not\models \psi \\
M,w,g \not\models \neg\psi & \text{iff} & M,w,g \models \psi \\
M,w,g \models ?\psi & \text{iff} & \text{neither } M,w,g \models \psi \text{ nor } M,w,g \not\models \psi \\
M,w,g \not\models ?\psi & \text{iff} & \text{either } M,w,g \models \psi \text{ or } M,w,g \not\models \psi \\
M,w,g \models \psi \wedge \chi & \text{iff} & M,w,g \models \psi \text{ and } M,w,g \models \chi \\
M,w,g \not\models \psi \wedge \chi & \text{iff} & M,w,g \not\models \psi \text{ or } M,w,g \not\models \chi \\
M,w,g \models \forall v\psi & \text{iff} & M,w,g' \models \psi \text{ for all } g' \text{ such that } g \overset{v}{\approx} g' \\
M,w,g \not\models \forall v\psi & \text{iff} & M,w,g' \not\models \psi \text{ for some } g' \text{ such that } g \overset{v}{\approx} g'
\end{array}
$$

Table 1: Satisfaction and violation conditions for the $\mathcal{TC}$ fragment of $\mathcal{MTC}$ (see Definition 2).

this view, a cause is typically neither a necessary condition for the effect (as typically some other event could, had it occurred, also have caused the effect), nor is a cause a sufficient condition for the effect (as typically there are many conditions which would, had they obtained or not obtained, have prevented the cause from having the effect). Consequently, Mackie defines a cause to be an insufficient but necessary part of an unnecessary but sufficient condition for its effect.

The formal definition of causation assumes the setting of a finite causal theory Θ, which may include definitions of the preconditions and postconditions of events. The definition begins with the requirement that a cause is part of an unnecessary but sufficient condition for its effect. Intuitively, the occurrence of event e at time t is a *sufficient cause* of effect ϕ if and only if e succeeds at t and ϕ is physically necessary given the postconditions of e at $t+1$. A physical necessity operator can be defined as follows: $\Box\phi \overset{def}{=} \bullet\phi \Rightarrow \phi$; thus $\Box\phi$ is true at a world w if and only if every accessible world above and below w is a ϕ-world. And, as Θ is assumed to be finite, the postcondition definitions in Θ can be represented by the $\mathcal{TC}$ sentence $\textit{Post}(\Theta) = \bigwedge\{\textit{Post}(e)(t) \equiv \phi \in \Theta\}$. Consequently, sufficient causation can be defined as follows:

$$
\begin{aligned}
&\textit{SCause}(\textit{Occ}(e)(t),\phi) \equiv \\
&\quad \textit{Succ}(e)(t) \wedge \Box(\textit{Post}(\Theta) \wedge \textit{Post}(e)(t+1) \rightarrow \phi)
\end{aligned} \tag{5}
$$

Turning now to the requirement that a cause is an insufficient but necessary condition for its effect, the occurrence of e at t is a *necessary cause* of effect ϕ if and only if e succeeds at t and no other event e' which occurs at t is a sufficient cause of ϕ:

$$
\begin{aligned}
&\textit{NCause}(\textit{Occ}(e)(t),\phi) \equiv \\
&\quad \textit{Succ}(e)(t) \wedge \bullet\exists e'(e \neq e' \wedge \textit{SCause}(\textit{Occ}(e')(t),\phi))
\end{aligned} \tag{6}
$$

Combining these two conditions gives the definition of causation:

$$
\begin{aligned}
&\textit{Cause}(\textit{Occ}(e)(t),\phi) \equiv \\
&\quad \textit{SCause}(\textit{Occ}(e)(t),\phi) \wedge \textit{NCause}(\textit{Occ}(e)(t),\phi)
\end{aligned} \tag{7}
$$

A law governing change and inertia can now be stated:

$$
\textit{Aff}(\ell)(t) \equiv \ell(t) \wedge \exists e Cause(\textit{Occ}(e)(t), \bullet\ell(t+1)) \tag{8}
$$

Thus the law states that Kleene literal ℓ is affected at time t if and only if ℓ is true at t and some event causes ℓ to have a different truth value at $t+1$. Less formally, the law states that nothing changes without a cause, or that every change has a cause.

A *causal theory* is any set of of sentences of $\mathcal{MTC}$ which contains the causal axioms $\Theta_C = \{(1),\ldots,(8)\}$.

The intended interpretation of causal theories is obtained by defining a suitable pragmatics for them. This is done by first defining the set of *causal models* for a causal theory Θ and then considering the consequences of Θ at a set of selected Θ-worlds in each of its causal models. The aim in defining the pragmatics is thus that the selected worlds in each causal model for Θ are just those worlds at which the axioms in Θ, especially the change and inertia axioms, are interpreted as intended.

$M, w, g =| \phi$*) according to the clauses given in tables 1 and 2.*

A formula ϕ *is* true *at a possible partial world* w *in an* $\mathcal{MTC}$*-model* M *(written* $M, w \models \phi$*) if* $M, w, g \models \phi$ *for all variable assignments* g*. A formula* ϕ *is* false *at* w *in* M *(written* $M, w =| \phi$*) if* $M, w, g =| \phi$ *for all variable assignments* g*.*

Proposition 3. *Let* M *be an* $\mathcal{MTC}$*-model containing world* w*, and let* ϕ *be a sentence of* $\mathcal{MTC}$*. Then either* $M, w \models \phi$ *or* $M, w \models \neg\phi$ *or* $M, w \models ?\phi$*.*

Thus, for a world w in an $\mathcal{MTC}$-model M, the truth and falsity conditions for sentences of $\mathcal{MTC}$ can be stated informally as follows. A sentence of the form $t < t'$ is true at w if and only if the time point denoted by t precedes that denoted by t', and is false at w otherwise. Similarly, the sentence $u = u'$ is true at w if u and u' denote the same object, and is false at w otherwise. An atomic sentence $r(u_1, \ldots, u_n)(t)$ it true at w if it is true that the relation holds between the objects denoted by $u_1, \ldots, u_n$ at time t, is false at w if it is false that the relation holds at t, and is undefined at w otherwise. Similarly, a higher-order atomic sentence $hr(e_1, \ldots, e_n, \ell_1, \ldots, \ell_m)(t)$ is true at w if it is true that the higher-order relation hr holds between event types $e_1, \ldots, e_n$ and the Kleene literals (see Section 5) $\ell_1, \ldots, \ell_m$ at time t, false at w if it is false that the relation holds at t, and is undefined at w otherwise. The clauses for negation, $\neg$, conjunction, $\wedge$, and the universal quantifier, $\forall$, follow those of Kleene. A sentence of the form $?\psi$ is true at w if the truth value of ψ is undefined at w, and is false in w otherwise. Finally, the clauses for $\Uparrow$ and $\Downarrow$ generalize those given in the previous section in a manner suggested by the semantics for classical counterfactuals given by Lewis [8] and Burgess [4].

The simpler truth conditions for counterfactuals which were given in the previous section can be stated formally using selection functions.

Definition 4. *Let* M *be a model,* ϕ *be a formula,* g *be a variable assignment, and* $[\![\phi]\!]^M_g$ *denote the set* $\{w : M, w, g \models \phi\}$ *of all worlds in* M *at which* g *satisfies* ϕ*. Then* $\uparrow(\phi, w, g) = \{w' : w \preceq w' \wedge w' \in [\![\phi]\!]^M_g \wedge \neg\exists w''(w \preceq w'' \prec w' \wedge w'' \in [\![\phi]\!]^M_g)\}$*; thus the function* $\uparrow$ *selects the closest worlds above* w *at which* g *satisfies* ϕ*. And* $\downarrow(\phi, w, g) = \{w' : w' \preceq w \wedge w' \in [\![\phi]\!]^M_g \wedge \neg\exists w''(w' \prec w'' \preceq w \wedge w'' \in [\![\phi]\!]^M_g)\}$*; thus the function* $\downarrow$ *selects the closest worlds below* w *at which* g *satisfies* ϕ*. Finally, for sets* S *and* T*,* $S \bullet T$ *("*S *overlaps* T*") if and only if* $S \cap T \neq \emptyset$*. Then the simplified satisfaction and violation conditions for complefactuals and contrafactuals are given in Table 3.*

Proposition 5. *If the limit assumption, LA, holds, then the simplified satisfaction and violation conditions for contrafactuals and complefactuals given in Table 3 are equivalent to the general satisfaction and violation conditions for contrafactuals and complefactuals given in Table 2.*

5 Actual events

A fairly comprehensive theory of common sense reasoning about actual events is developed in [2]; according to which events are defeasible, they may be non-deterministic, they may have context-dependent effects, and they may occur simultaneously. This section shows how the sub-theory of "primary" events can be embedded in $\mathcal{MTC}$; the inclusion of the rest of the theory being straightforward.

Primary events can be thought of as defeasible STRIPS events. Thus they are defined by specifying their preconditions and their postconditions; examples of these are (10) and (11) in the following section. The axiom of change then states that if event e occurs at time t and the preconditions of e are true at t and it is not true that e is qualified at t then the postconditions of e are true at $t + 1$:

$$\mathit{Pre}(e)(t) \wedge \mathit{Occ}(e)(t) \wedge \bullet\mathit{Qual}(e)(t) \rightarrow \mathit{Post}(e)(t+1) \tag{1}$$

Intuitively, e is qualified at t if there is some reason why e should not succeed at t. The intention is to use this axiom positively whenever possible: given $\mathit{Pre}(e)(t)$ and $\mathit{Occ}(e)(t)$, $?\mathit{Qual}(e)(t)$ should be assumed and the axiom used to conclude $\mathit{Post}(e)(t+1)$, if doing so is consistent. Thus on the intended interpretation of the axiom events normally succeed if their preconditions are true when they occur. Qualifications apply only to events which would otherwise succeed:

$$\mathit{Qual}(e)(t) \rightarrow \mathit{Pre}(e)(t) \wedge \mathit{Occ}(e)(t) \tag{2}$$

And the distinction between the occurrence of an event and its success is highlighted by the following axiom:

$$\mathit{Succ}(e)(t) \equiv \mathit{Pre}(e)(t) \wedge \mathit{Occ}(e)(t) \wedge \bullet\mathit{Qual}(e)(t) \tag{3}$$

Inertia is represented by means of a common sense inertia axiom. Intuitively, if an atom of the form $r(u_1, \ldots, u_n)(t)$ is true, and there is no reason to doubt that the relation $r(u_1, \ldots, u_n)$ persists, we should conclude that it does so; that is, that $r(u_1, \ldots, u_n)(t+1)$ is true. Similarly, negated atoms of this form should persist by default. In order to formalize this, the non-temporal component $r(u_1, \ldots, u_n)$ of an atom $r(u_1, \ldots, u_n)(t)$ is called a *Kleene atom*, and a *Kleene literal* is either a Kleene atom or its negation. Then, for Kleene literal ℓ and time t, $\mathit{Aff}(\ell)(t)$ states that ℓ is affected at t; that is, that there is reason to doubt that that the truth value of ℓ persists beyond t. The inertia axiom is thus as follows:

$$\ell(t) \wedge \bullet\mathit{Aff}(\ell)(t) \rightarrow \ell(t+1) \tag{4}$$

Thus the axiom states that if the Kleene literal ℓ is true at time t and it is not true that ℓ is affected at t then ℓ remains true at $t + 1$. The intention is that the axiom should be used positively whenever possible: given $\ell(t)$, $?\mathit{Aff}(\ell)(t)$ should be assumed and the axiom used to conclude $\ell(t+1)$ if doing so is consistent.

We are now in a position to give a definition of direct causation. The relation *Cause* holds between sentences of the form $\mathit{Occ}(e)(t)$ and sentences of $\mathcal{MTC}$. The intended reading of a sentence of the form $\mathit{Cause}(\mathit{Occ}(e)(t), \phi)$ is thus that the occurrence of event e at time t is the (direct) cause of ϕ. The definition can be seen as a formalization of Mackie's characterization of causes as INUS conditions [10]. According to

conditional of this kind will be called a *contrafactual*, as the truth of its antecedent contradicts what is true at w, and will be written $\phi \Downarrow \psi$. The contrafactual $\phi \Downarrow \psi$ should thus be false at w if at least one of the closest worlds below w where where ϕ is not false is a $\neg(\phi \Uparrow \psi)$-world.

Consequently a counterfactual $\phi \Rightarrow \psi$ should be true at w if either the complefactual $\phi \Uparrow \psi$ is true at w or the contrafactual $\phi \Downarrow \psi$ is true at w, and $\phi \Rightarrow \psi$ should be false at w if either $\phi \Uparrow \psi$ or $\phi \Downarrow \psi$ is false at w.

The semantics for $\Uparrow$ may give unintuitive results if there are infinitely descending sequences of ϕ-worlds above w, as there may then be ϕ-worlds above w, but no *closest* ϕ-worlds above w. Similarly the semantics for $\Downarrow$ may give unintuitive results if there are infinitely ascending sequences of ϕ-worlds below w. In many practical applications there are no such sequences, so the following counterpart of Lewis' Limit Assumption holds:

> (LA) For every ϕ-world w' such that $w \preceq w'$ there is a world w'' such that $w \preceq w'' \preceq w'$ which is a closest ϕ-world above w. For every ϕ-world w' such that $w' \preceq w$ there is a world w'' such that $w' \preceq w'' \preceq w$ which is a closest ϕ-world below w.

More general semantics are given in Section 4 which do not depend on condition LA, but which reduce to the simpler closest-world semantics given here when LA does hold.

4 The modal temporal calculus

In order to represent reasoning about events at each possible partial world, the modal language of the previous section is now combined with the Temporal Calculus, or $\mathcal{TC}$ [2], resulting in the modal temporal calculus, or $\mathcal{MTC}$.

Recall the practical, resource-bounded, interpretation of $\mathcal{TC}$. Thus a sentence ϕ is true (false) if the truth (falsity) of ϕ is relevant and can be established given the limited resources available, and is undefined otherwise. Recall also that the undefined operator, ?, significantly increases the expressiveness of the language, as illustrated by the following definitions:

$$\begin{array}{rcl} \circ\phi & \stackrel{def}{=} & ?\phi \vee \phi \\ \bullet\phi & \stackrel{def}{=} & ?\phi \vee \neg\phi \\ !\phi & \stackrel{def}{=} & \neg?\phi \\ \phi \rightarrow \psi & \stackrel{def}{=} & \bullet\phi \vee \neg\bullet\psi \\ \phi \equiv \psi & \stackrel{def}{=} & (\neg\bullet\phi \wedge \neg\bullet\psi) \vee (\neg\circ\phi \wedge \neg\circ\psi) \vee (?\phi \wedge ?\psi) \end{array}$$

Thus, for sentence ϕ, $\circ\phi$ states that ϕ is not false (that ϕ is either undefined or true), $\bullet\phi$ states that ϕ is not true (that ϕ is either undefined or false), and $!\phi$ states that the truth value of ϕ is defined (is either true or false). For sentences ϕ and ψ, the conditional $\phi \rightarrow \psi$ is false if ϕ is true and ψ is not, and is true otherwise. Thus, in keeping with the resource-bounded interpretation, the conditional can be thought of as expressing a constraint which must be met if the antecedent is true, but which can otherwise be ignored. Finally, for sentences ϕ and ψ, the equivalence $\phi \equiv \psi$ is true if ϕ and ψ have the same truth value (true, false, undefined).

The language $\mathcal{MTC}$ is obtained by adding the binary connectives $\Uparrow$, $\Downarrow$ and $\Rightarrow$ to $\mathcal{TC}$.

The semantics of $\mathcal{MTC}$ is given by combining the semantics of the previous section with the semantics of $\mathcal{TC}$.

Definition 1. *A* model *for $\mathcal{MTC}$ is a structure:*

$$M = \langle \mathcal{W}, \prec, \mathcal{D}, \mathcal{E}, \mathcal{T}, \prec_T, \mathcal{F}, \mathcal{R}, \mathcal{HR}, \mathcal{V} \rangle,$$

where:

- *$\mathcal{W}$, $\mathcal{D}$, $\mathcal{E}$ and $\mathcal{T}$ are mutually disjoint non-empty sets,*
- *$\prec$ is a strict partial order on $\mathcal{W}$; thus, $\prec$ is a binary relation on $\prec$ which is irreflexive and transitive,*
- *$\prec_T$ is a binary relation on $\mathcal{T}$ which is discrete and linear,*
- *$\mathcal{F} = \langle \mathcal{F}_D, \mathcal{F}_T, \mathcal{F}_E \rangle$, where $\mathcal{F}_S$ is a set of n-ary functions of type $\mathcal{S}^n \rightarrow \mathcal{S}$, for $n \geq 1$ and $\langle S, \mathcal{S} \rangle \in \{\langle D, \mathcal{D} \rangle, \langle T, \mathcal{T} \rangle, \langle E, \mathcal{E} \rangle\}$,*
- *$\mathcal{R}$ is a set of partial n-ary functions of type $\mathcal{D}^n \rightarrow \{true, false\}$ for $n \geq 0$,*
- *$\mathcal{HR}$ is a set of partial $n+m$-ary functions of type $\mathcal{E}^n \times L^m \rightarrow \{true, false\}$ for $n + m \geq 1$,*
- *$\mathcal{V} = \langle \mathcal{V}_D, \mathcal{V}_T, \mathcal{V}_E, \mathcal{V}_L, \mathcal{V}_{F_D}, \mathcal{V}_{F_T}, \mathcal{V}_{F_E}, \mathcal{V}_R, \mathcal{V}_{HR}, \rangle$ is an interpretation function such that:*
  - *$\mathcal{V}_S : S \rightarrow \mathcal{S}$ for $\langle S, \mathcal{S} \rangle \in \{\langle D, \mathcal{D} \rangle, \langle T, \mathcal{T} \rangle, \langle E, \mathcal{E} \rangle\}$,*
  - *$\mathcal{V}_L : L \rightarrow L$ is the identity function,*
  - *$\mathcal{V}_{F_S} : F_S \rightarrow \mathcal{F}_S$ for $\langle S, \mathcal{S} \rangle \in \{\langle D, \mathcal{D} \rangle, \langle T, \mathcal{T} \rangle, \langle E, \mathcal{E} \rangle\}$,*
  - *$\mathcal{V}_R : R \times \mathcal{W} \times \mathcal{T} \rightarrow \mathcal{R}$, and*
  - *$\mathcal{V}_{HR} : HR \times \mathcal{W} \times \mathcal{T} \rightarrow \mathcal{HR}$.*

Intuitively, $\mathcal{W}$ is a set of partial possible worlds and $\prec$ is the closeness relation on worlds; thus if $w \prec w' \prec w''$, then w' is, in the appropriate sense, closer to w than w'' is. The reflexive closure of $\prec$ is defined in the usual way; thus $w \preceq w'$ if and only if either $w \prec w'$, or $w \in \mathcal{W}$ and $w = w'$. The members of $\mathcal{D}$ should be thought of as material objects, while the members of $\mathcal{E}$ should be thought of as event types. Note that, for simplicity, the set of domain objects and event types is fixed across worlds. Time is represented by the temporal frame $\langle \mathcal{T}, \prec_T \rangle$, where $\mathcal{T}$ is a set of time points and $\prec_T$ is the before-after relation. Note that all worlds share a common time. The interpretation function $\mathcal{V}$ is defined so as to allow the extension of first-order and higher-order relations to vary across worlds and times while, again for simplicity, keeping the denotation of constants and functions fixed across worlds and times.

Terms are interpreted in the standard way with the exception of Kleene literals (see Section 5), which are "interpreted" as themselves. The formal definitions follow those for $\mathcal{TC}$ given in [2] and so will not be repeated here.

The truth and falsity of sentences at each world is defined by means of the intermediary notions of the satisfaction and violation of formulas at that world.

Definition 2. *Let $M = \langle \mathcal{W}, \prec, \mathcal{D}, \mathcal{E}, \mathcal{T}, \prec_T, \mathcal{F}, \mathcal{R}, \mathcal{HR}, \mathcal{V} \rangle$ be an $\mathcal{MTC}$-model, g be a variable assignment for M, and ϕ be an $\mathcal{MTC}$-formula. Then g* satisfies *ϕ at a world w in M (written $M, w, g \models \phi$) or* violates *ϕ at w in M (written*

is a ϕ-world w' which is accessible from w and every ϕ-world w' which is at least as similar as w' is to w is also a ψ-world. Once again, a number of semantic conditions are imposed in order to ensure the semantic integrity of the pragmatic parameter. Thus each comparative similarity relation $\preceq_w$ is required to be a weak order on the set of accessible worlds W_w (a linear order on equivalence classes of W_w) which is centered on w (that is, $w \in W_w$ and, for any $w' \in W_w$, $w \preceq_w w'$).

However, the vagueness which gives these analyses their strength is also their weakness. In order to evaluate a counterfactual it is necessary to choose an appropriate value for the pragmatic parameter. But while the semantic conditions constrain the choice of this value, they do not determine it. Thus Stalnaker defines the pragmatic problem of counterfactuals to be that of finding and defending criteria for choosing appropriate values for the pragmatic parameter.

This has typically been done informally. For example, Lewis notes that if in choosing a similarity relation "we try too hard for exact similarity in one respect, we will get excessive differences in some other respects"; for example, a world in which kangaroos have no tails and everything else is as it actually is, is a world with physics and genetics which are very different from those of the actual world [8, p. 9]. Consequently, "respects of similarity and difference trade off", [8, p. 9], and "[o]verall similarity among worlds is some sort of resultant of similarities and differences of many different kinds" [9, p. 465].

Lewis counts it as a virtue of his analysis that the comparative similarity relation is not specified more formally: "I have not said what system of weights or priorities should be used to squeeze these [similarities and differences] down into a single relation of overall similarity. . . . Counterfactuals are both vague and various. Different resolutions of the vagueness of overall similarity are appropriate in different contexts" [9, p. 465].

However, the vagueness of the notion comparative similarity has proved problematic, as the discussion in Example 11 in Section 6 shows. Moreover if counterfactuals are to be of use in Artificial Intelligence, then it is necessary to provide formal pragmatics for them. This will be done for causal counterfactuals. But, before doing so, an appropriate partial semantics is defined for them.

3 Complefactuals and contrafactuals

The concept of causation arises from the need to reason about events and their effects on the basis of incomplete, or partial, information, and consequently any formal treatment of causation should reflect this. However classical logic assumes complete, or total, information, and consequently any attempt to use it to represent causality, and indeed common sense reasoning generally, involves some means for introducing partiality. Thus, for example, possible worlds are total, in the sense that the truth value of each proposition is decided (as either true or false) at that world. Partiality can be introduced by considering what is true (false) in a set of possible worlds, however it is desirable to represent partiality in a direct and less artificial way. Thus we begin with the idea of a *possible partial world*.

A possible partial world w can be thought of as a set of classical possible worlds; some sentences may be true at w, others false, and yet others may be undefined (neither true nor false). Accordingly the semantics of the language of each possible partial world is that proposed by Kleene [7], which agrees with the classical truth-functional semantics wherever possible. Thus, an atomic sentence p may be either true, false or undefined at w; a sentence $\neg\phi$ is true at w if ϕ is false at w, false if ϕ is true at w, and is undefined otherwise; and a sentence $\phi \wedge \psi$ is true at w if ϕ and ψ are both true, false at w if either is false, and is undefined at w otherwise. The connectives $\vee$ and $\supset$ can be defined in the same way as their classical counterparts; thus, for example, $\phi \vee \psi$ is defined as $\neg(\neg\phi \wedge \neg\psi)$. For the sake of convenience, possible partial worlds will often be referred to simply as "worlds", and a world at which sentence ϕ is true will be referred to as a "ϕ-world".

A possible partial worlds model is a triple $M = \langle \mathcal{W}, \prec, \mathcal{V} \rangle$, where $\mathcal{W}$ is a nonempty set of possible partial worlds, $\prec$ is a binary relation on $\mathcal{W}$, and $\mathcal{V}$ is a function with domain $\mathcal{W}$. For each $w \in \mathcal{W}$, $\mathcal{V}_w$ is a partial function which assigns at most one of the values *true* or *false* to each atomic sentence p. In order to represent the growth of information, $\prec$ can be thought of as an information ordering on worlds. Thus a possible partial worlds model M is said to be an *information model* if $\prec$ is a strict partial order on $\mathcal{W}$ (if $\prec$ is irreflexive and transitive) and $\mathcal{V}$ satisfies the following condition:

$$\text{(Persistence)} \quad \text{If } w \prec w' \text{ then } \mathcal{V}_w \subset \mathcal{V}_{w'};$$

where $\mathcal{V}_w \subset \mathcal{V}_{w'}$ if $\mathcal{V}_{w'}$ extends $\mathcal{V}_w$; that is, if $\mathcal{V}_{w'}(p) = \mathcal{V}_w(p)$ whenever $\mathcal{V}_w(p) = \textit{true}$ or $\mathcal{V}_w(p) = \textit{false}$, and there is at least one p such that $\mathcal{V}_w(p)$ is undefined and $\mathcal{V}_{w'}(p)$ is not. Thus if $w \prec w'$ and the atomic sentence p is true (or false) at w, then the truth value of p persists at w'. It follows that if $w \prec w'$ and the sentence ϕ is true (false) at w, then its truth value persists at w'. So if $w \prec w'$, then w' contains more information than w; that is, w' is a better approximation of a classical possible world than w is.

The question now arises: what semantics can be given for counterfactuals in information models? By analogy with the classical analysis, a counterfactual $\phi \Rightarrow \psi$ should be true if the truth of ϕ and non-truth of ψ is, in some sense, a remoter possibility than the truth of $\phi \wedge \psi$. However, in view of persistence, it seems that there are two possibilities.

If ϕ is not false at a world w, then the counterfactual $\phi \Rightarrow \psi$ should be true at w just in case all of the closest ϕ-worlds above ϕ are also ψ-worlds; where w' is a *closest ϕ-world above* w if $w \preceq w'$, ϕ is true at w', and there is no other ϕ-world w'' such that $w \preceq w'' \preceq w'$; where $w \preceq w'$ if $w \prec w'$, or $w = w'$ and $w \in \mathcal{W}$. A conditional of this kind will be called a *complefactual*, as the truth of its antecedent complements what is true at w, and will be written $\phi \Uparrow \psi$. The complefactual $\phi \Uparrow \psi$ should thus be false at w if at least one of the closest ϕ-worlds above w is a $\neg\psi$-world.

Alternatively, if ϕ is false at w, then the counterfactual $\phi \Rightarrow \psi$ should be true at w just in case all of the closest worlds below w where ϕ is not false are all $\phi \Uparrow \psi$-worlds; where w' is a *closest ϕ-world below* w if $w' \preceq w$, ϕ is true at w', and there is no other ϕ-world w'' such that $w' \preceq w'' \preceq w$. A

Causal Counterfactuals

John Bell
Applied Logic Group
Department of Computer Science
Queen Mary, University of London
London E1 4NS
jb@dcs.qmw.ac.uk

Abstract

The formal possible-worlds analysis of counterfactuals has tended to concentrate on their semantics and logic, with their pragmatics being given informally. However, if counterfactuals are to be of use in Artificial Intelligence, it is necessary to provide formal pragmatics for them. This is done in this paper by combining work on the representation of common sense reasoning about events with an appropriate semantics for counterfactuals. The resulting combination provides a unified framework for formal reasoning about actual and counterfactual events.

1 Introduction

Counterfactuals play an essential role in practical reasoning. Intelligent agents need to be able to reason counterfactually about the consequences of actions and events. For example, a planning agent needs to be able to reason that a plan is likely to achieve a goal if it is executed, and if the plan has to be revised during execution, that the revised plan will probably succeed. An agent which can reason counterfactually in this way can also benefit from the ability to form contingency plans, to reason that if a plan were to go awry at some stage of its execution, then an alternative plan would be appropriate. An agent of this kind can also benefit from hindsight. If a plan has failed, the agent can learn from this experience by considering which alternative plans would have succeeded.

The importance of counterfactuals in Artificial Intelligence has long been recognized; for example in [11]. Recent work includes Pearl's probabilistic theory [12; 13], and Costello and McCarthy's Cartesian counterfactuals [5].

This paper presents a theory of *causal counterfactuals*, which combines an appropriate semantics for counterfactuals with the theory of common sense reasoning about actual events which is developed in [2]. The semantics of the theory can be seen as a development of the classical possible-worlds semantics of Stalnaker, Thomason, and Lewis [8; 15; 16], which are outlined in the next section. The new semantics are then presented informally in Section 3 in the general setting of information models [17]. In Section 4 the semantics is combined with the language used for reasoning about actual events in [2], resulting in a language called the Modal Temporal Calculus, or $\mathcal{MTC}$. In Section 5 the common sense theory of events developed in [2] is embedded in $\mathcal{MTC}$, and an appropriate formal pragmatics is given for the new setting. The pragmatics is appropriate for reasoning about actual events, and, as Section 6 suggests, for causal counterfactuals and reasoning about counterfactual events.

2 Classical semantic theories

Counterfactuals are notoriously vague and context-dependent. Nevertheless, Stalnaker [15] argues that it is possible to give a semantic analysis of them which includes what might be called a pragmatic parameter. Thus his truth conditions include a selection function on possible worlds which, for each possible world w and proposition ϕ selects the closest ϕ-world to w; where a ϕ-world is a world at which ϕ is true. Then a conditional sentence $\phi > \psi$ is true at a possible world w if and only if the selected ϕ-world is also a ψ-world. In order to ensure that an appropriate world is selected, Stalnaker imposes a number of general, semantic, conditions on the selection function; for example, if w is a ϕ-world, then it should be selected as the closest ϕ-world to itself.

Stalnaker argues that the advantage of such an analysis is that it is possible to draw a clear distinction between the semantics of counterfactuals and their pragmatics. The semantics for counterfactuals brings out the common structure of their truth conditions by giving the counterfactual connective a single meaning and making their pragmatics a parameter of the interpretation. Consequently it is possible to define semantical notions such as validity and consequence, and to give sound and complete axiomatizations for counterfactuals; as is done by Stalnaker and Thomason in [16].

Lewis [8] argues that there is typically not a single closest ϕ-world to a given world w, but rather a set of such worlds. Consequently he generalizes Stalnaker's analysis by having the selection function return the set of closest ϕ-worlds to w. He also gives an alternative semantics, in which worlds are ordered according to their comparative overall similarity to the actual world. Thus it is assumed that for each possible world w the set W_w of all worlds which are accessible from w can be ordered by the comparative similarity relation $\preceq_w$, where $w' \preceq_w w''$ holds if and only if w' is at least as similar to w as w'' is. The counterfactual $\phi \Box\!\!\rightarrow \psi$ is then true at w if and only if either no ϕ-worlds are accessible from w, or there