Causation and Causal Conditionals*

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

Abstract

Causation is defined recursively: event e is the cause of condition ϕ in context c iff e is the only sufficient cause of ϕ in c, and removing e from c either removes ϕ from c or results in some other event causing ϕ. A logical language is then defined, in which it is possible to represent and reason about actual and counterfactual events in evolving partial contexts. Axiomatic theories of events and causation are given, and a formal pragmatics is defined, making it possible to reason formally about particular cases. By way of illustration, examples involving preemption and trumping preemption are given.

Introduction

Lewis (1973) observes that Hume defined causation "twice over":

> [W]e may define a cause to be *an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second*. Or, in other words *where, if the first object had not been, the second never had existed*. (Hume 1975, §VII, Part II)

Hume's first definition characterizes causes as being sufficient for their effects, his second as being necessary for them.

In (Bell 2003), I suggest combining contextual necessity and sufficiency. Event e is sufficient for condition ϕ in context c iff e succeeds in c and ϕ is among e's effects. Event e is necessary for ϕ in c iff removing e from c also removes ϕ from c. Then e is the (direct) cause of ϕ iff e is both necessary and sufficient for ϕ in context c. I then develop a formal theory of contextual sufficiency, in which it is reduced to the elements of a common sense theory of natural (or defeasible) events. However, while the theory captures many aspects of contextual sufficiency, it only captures contextual necessity in the special case in which the effects of an event change the context. The theory is thus too liberal in attributing causal status to events. For example, whitewashing an already white wall may be contextually sufficient for the wall's being white, but is not contextually necessary for it, and so should not be considered to be its cause. Consequently this paper is devoted to extending the theory to include a better account of contextual necessity.

Lewis (1973) formalizes Hume's necessity definition using his possible-worlds theory of counterfactuals. The occurrence of event e, $O(e)$, being counterfactually dependent on the occurrence of event c, $O(c)$, iff $O(c)$ and $O(e)$ are both true, and the counterfactual $\neg O(c) \,\Box\!\!\rightarrow \neg O(e)$ ("If c had not occurred, then e would not have occurred") is true.

Lewis's theory has been much discussed; see, for example, (Lewis 1986, Ch. 21). A serious drawback of the theory is the lack of a formal pragmatics; how, exactly, are particular counterfactuals evaluated? A further problem is posed by the phenomenon of trumping preemption (Schaffer 2000). Preemtion occurs when the effects of one event prevent another event from having the same (or similar) effects. For example (Wright 1988), suppose that two fires, A and B advance toward a house from opposite directions, and that A arrives first and burns the house down. Then fire A preempts fire B from destroying the house, and is considered to be the cause of the house's destruction, despite it being granted that if A had not destroyed the house, then B would have done so. Examples such as this can be dealt with by distinguishing between two events, the house's actual destruction by A at time t, and its preempted later destruction by B at time $t' > t$, then A causes the house's destruction at t, as otherwise B would still not reach the house until t'. However, distinctions such as this don't work when one event *trumps* another. Lewis (2000) gives the following example; suggested by Bas van Fraassen. Suppose that a sergeant and a major simultaneously order their soldiers to advance, and that the soldiers do so. Their advance is redundantly caused, since either order would, on its own, have been sufficient. However, the redundancy is asymmetric, since the soldiers obey the senior officer. The soldiers advance because the major orders them to, not because the sergeant does. The major's order trumps the sergeant's. Consequently Lewis (2000) proposes a revised theory of causation as influence. Causes can, he suggests, be distinguished by looking at the pattern of counterfactual dependence of alterations of the effect upon alterations of the cause. Thus, in a case of trumping, the real cause can be distinguished from an event it trumps by the fact that altering the cause slightly alters the effect slightly, whereas altering the trumped event slightly

*I am grateful to the reviewers and participants of *Common Sense 2001* and *Context 2003*, the *KR 2004* reviewers, and Wilfrid Hodges for helpful discussions and comments.

does not alter the effect.

However, it seems to me that, given an account of contextual sufficiency, Lewis's original counterfactual-dependency account can be retained, but with a refinement which makes the definition of causation recursive:

> Event e is the cause of condition ϕ in context c iff e is the only sufficient cause of ϕ in c, and removing e from c either removes ϕ from c or results in some other event causing ϕ.

Thus, on the parade ground, the sergeant's order is inoperative, because the soldiers are obliged to obey an order from a more senior officer, and consequently the major's order is the only sufficient cause of the soldiers' advance. However, if the major's order had not been given (or was not heard, etc.), then the sergeant's order would have caused the advance; as it would then be both contextually sufficient and contextually necessary (being the only other order issued). The recursive nature of the definition makes it capable of dealing with serial trumping; as might occur, for example, if the whole chain of command gave the same order simultaneously.

A formal version of the proposed definition is developed in this paper. The formal sufficiency theory is defined in a three-valued language of events, and is logico-pragmatic. Models of the language are partially ordered according to their chronological minimality. Roughly speaking, a model is preferred to another if they agree up to some time point (if they represent a common history up to that point) and the preferred model contains less information at that point (for example, if the history it represents is less eventful at that point). The intended interpretation of a given theory is then obtained by focussing on what is true in all of its most preferred models. The key insight of the extension given here (first essayed in (Bell 2001a)) is that the pragmatics of the sufficiency theory can be used as the basis for a formal theory of causal counterfactuals. Semantically, models of the language of events can be thought of as possible partial worlds, and the preference relation on them can be thought of as an accessibility relation. The accessibility relation also provides the basis of a formal pragmatics, as it orders worlds (models) according to their comparative chronological similarity; worlds above (below) a world in the ordering represent alternative histories which differ at some point because of the addition (deletion) of certain facts or events.

However, in order to make this idea work, it is necessary to develop an appropriate semantics for counterfactuals in partial orders of the kind envisaged. It is also necessary to refine the pragmatics, essentially by reordering the worlds in a given frame so that the histories of related worlds represent genuine alternative histories.

An appropriate semantics for conditionals is given in the next section. The theory of events on which the sufficiency theory is based is then recalled in Section 3. The definition of causation is given in Section 4, and its formal pragmatics is defined in Section 5. In order to illustrate the theory, formal versions of the two-fires and military-trumping examples are given in Section 6.

The Causal Language $\mathcal{CL}$

The causal language $\mathcal{CL}$ has been developed in order to be able to represent and reason about actual and counterfactual events in evolving partial contexts. This section begins with an informal introduction to $\mathcal{CL}$. The formal syntax and semantics are then given.

We begin with the language of events, $\mathcal{EL}$. $\mathcal{EL}$ is based on Kleene's (1952) strong three-valued language which provides a means for reasoning demi-classically with partial information and classically with complete information. Accordingly, the truth conditions for the propositional operators return a Boolean truth value wherever possible. An atomic sentence p may be either true, false or undefined; the sentence $\neg\phi$ is true if ϕ is false, false if ϕ is true, and is undefined otherwise; and the sentence $\phi \wedge \psi$ is true if ϕ and ψ are both true, false if either is false, and is undefined otherwise. Further operators, such as inclusive disjunction, can be defined as in classical logic: thus $\phi \vee \psi =_{\mathrm{Df}} \neg(\neg\phi \wedge \neg\psi)$. The first-order extension is straightforward. Atomic sentences may be true, false, or undefined; a universal sentence $\forall x\phi$ is true if ϕ is true for all assignments to x, false if ϕ is false for one such assignment, and is undefined otherwise; and the existential quanitifier $\exists$ is defined as in classical logic.

In order to represent and reason about partiality, the *undefined* operator 'U' is added to Kleene's language. The sentence $\mathsf{U}\phi$ is true if ϕ is undefined (is neither true nor false), and is false otherwise. This operator is used to define the classically-valued operators T, F, $\rightarrow$ and $\equiv$ as follows:

$$\mathsf{T}\phi =_{\mathrm{Df}} \neg(\mathsf{U}\phi \vee \neg\phi), \quad \mathsf{F}\phi =_{\mathrm{Df}} \neg(\mathsf{U}\phi \vee \phi),$$
$$\phi \rightarrow \psi =_{\mathrm{Df}} \neg\mathsf{T}\phi \vee \mathsf{T}\psi,$$
$$\phi \equiv \psi =_{\mathrm{Df}} (\mathsf{T}\phi \wedge \mathsf{T}\psi) \vee (\mathsf{F}\phi \wedge \mathsf{F}\psi) \vee (\mathsf{U}\phi \wedge \mathsf{U}\psi).$$

Thus, for sentences ϕ and ψ: $\mathsf{T}\phi$ is true if ϕ is true, and is false otherwise; $\mathsf{F}\phi$ is true if ϕ is false, and is false otherwise; and $\phi \rightarrow \psi$ is true if ψ is true or ϕ is not, and is false otherwise; and $\phi \equiv \psi$ is true if ϕ and ψ have the same truth value, and is false otherwise.

In order to represent time, time points are added as a second sort. For simplicity, time is assumed to be discrete and linear, and relations between time points (identity and precedence) are defined classically. The time-dependent nature of facts is then represented by adding a temporal index to atoms of the underlying language. Thus a *domain atom* is an atom of the form $r(u_1, \ldots, u_n)(t)$, where the u_i are terms denoting objects in the domain, and term t denotes a time point. Intuitively, a domain atom $r(u_1, \ldots, u_n)(t)$ states that the relation r holds between the objects $u_1, \ldots, u_n$ at time (point) t, that the fact $r(u_1, \ldots, u_n)$ is true at t. Formally, *facts* are defined to be the atemporal components of temporal literals; thus if $\alpha(t)$ is a domain atom, then α and $\neg\alpha$ are both facts.

In order to reason about inertia (the persistence of facts over time) facts are added as a third sort and higher-order quantification over them is introduced.

Finally, events are added as a fourth sort. For example, an *occurs atom* is an atom of the form $\mathit{Occ}(e)(t)$, which states that event e (or, more precisely, a token of event type e) occurs at time t. More generally, an *event atom* is an atom of the form $r(e_1, \ldots, e_n)(t)$, where each e_i is an event term.

In order to define causation, $\mathcal{EL}$ is extended to the full causal language $\mathcal{CL}$ by adding quantification over formulas of $\mathcal{EL}$, and the modal operators $\Box$, $\Uparrow$, $\Downarrow$, and $\Rightarrow$.

The semantics of the modal operators is given by introducing partial worlds frames, each consisting of a set of possible partial worlds, $\mathcal{W}$, and a partial order $\prec$ on $\mathcal{W}$ which represents accessibility among possible partial worlds. Possible partial worlds are like the possible worlds of normal classical modal logics, except that the truth values of some atomic propositions may be undefined at them; thus a possible partial world may be thought of as a set of (classical) possible worlds, or, more naturally, as a partially specified (classical) possible world. In the sequel possible partial worlds will be referred to simply as "worlds", and a world at which sentence ϕ (set of sentences Θ) is true will be referred to as a ϕ-world (a Θ-world). In this setting, the extensional event language $\mathcal{EL}$ is used to describe individual worlds, while the intensional (modal) operators of $\mathcal{CL}$ are used to refer across worlds.

For a given set of worlds $\mathcal{W}$, $\Box\phi$ states that ϕ is true at all worlds in $\mathcal{W}$, that ϕ is logically necessary given the truths of $\mathcal{W}$. When, as in the intended use, these truths consist of a set of physical laws, $\Box\phi$ can be understood as stating that ϕ is physically necessary

The semantics of the conditional operators $\Uparrow$, $\Downarrow$, and $\Rightarrow$, is given in terms of the set of closest antecedent worlds to any given world. As usual, define $w \preceq w'$ iff either $w \prec w'$, or $w = w'$ and $w \in \mathcal{W}$. Then, for worlds $w, w' \in \mathcal{W}$, w' is a *closest ϕ-world above* w iff $w \preceq w'$, w' is a ϕ-world, and there is no ϕ-world w'' such that $w \preceq w'' \prec w'$. Similarly, for worlds $w, w' \in \mathcal{W}$, w' is a *closest ϕ-world below* w iff $w' \preceq w$, w' is a ϕ-world, and there is no ϕ-world w'' such that $w' \prec w'' \preceq w$. Thus if w is a ϕ-world, then it is also the closest ϕ-world above and below itself.

It is assumed that the accessibility relation $\prec$ reflects some form of *vertical persistence of information*; that is, if $w \prec w'$, then w' contains more information, in some sense, than w. In particular, the order $\prec_C^\Theta$ defined in Section 5 is based on monotonically increasing support sets; that is conditions which, together with the laws of a common causal theory, determine the evolution of worlds. Thus if $w \prec_C^\Theta w'$, then the history of w' is richer than (is perhaps better determined, or perhaps more eventful than) that of w, and so requires more support than w. Note that the particular details of vertical persistence may vary according to the application, it is thus a pragmatic condition, rather than a semantic one.

A *complefactual* is a sentence of the form $\phi \Uparrow \psi$. Intuitively $\phi \Uparrow \psi$ is true at world w if ψ is true at the closest worlds above w at which the vertically persistent information at w is complemented by ϕ. Accordingly, the complefactual $\phi \Uparrow \psi$ is applicable at world w if ϕ is not false at w, in which case it is true if the closest ϕ-worlds above w are all ψ-worlds, and false if one of these worlds is a $\neg\mathsf{T}\psi$-world. The evaluation of a complefactual can thus be thought of as involving an AGM expansion operation (Gärdenfors 1988). Given that the complefactual $\phi \Uparrow \psi$ is applicable at w, it is true at w iff all minimal expansions of the persistent information at w which make ϕ true also make ψ true.

A *contrafactual* is a sentence of the form $\phi \Downarrow \psi$. Intuitively $\phi \Downarrow \psi$ is true at world w if ϕ is false at w and the complefactual $\phi \Uparrow \psi$ is true at the closest worlds below w at which ϕ is not false. Accordingly the contrafactual $\phi \Downarrow \psi$ is applicable at world w if ϕ is false at w, in which case it is true if the closest $\neg\mathsf{F}\phi$-worlds below w are all $\phi \Uparrow \psi$-worlds, and false if one of these worlds is a $\neg\mathsf{T}(\phi \Uparrow \psi)$-world. The evaluation of a contrafactual can thus be thought of as involving an AGM revision operation (consisting of a contraction operation followed by an expansion operation). Given that the contrafactual $\phi \Downarrow \psi$ is applicable at w, it is true at w iff all minimal contractions of the persistent information at w which make $\neg\mathsf{F}\phi$ true also make $\phi \Uparrow \psi$ true.

By analogy with the classical analysis, a *counterfactual* sentence $\phi \Rightarrow \psi$ should be true if the truth of $\phi \wedge \neg\mathsf{T}\psi$ is, in some sense, a remoter possibility than the truth of $\phi \wedge \psi$. At a world w in a partial order with vertically persistent information there are two possibilities. If ϕ is not false at w, then the counterfactual $\phi \Rightarrow \psi$ should be true (false) at w iff the complefactual $\phi \Uparrow \psi$ is true (false) at w. Alternatively, if ϕ is false at w, then the counterfactual $\phi \Rightarrow \psi$ should be true (false) at w just in case the contrafactual $\phi \Downarrow \psi$ is true (false) at w. 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.

A formal account of $\mathcal{CL}$ is now given. The five sorts of $\mathcal{CL}$ are identified by the following letters: D for domain objects, T for time points, and E for events, F for facts, and Φ for $\mathcal{EL}$-formulas.

Definition 1. *The vocabulary of $\mathcal{CL}$ consists of the symbols '$<$', '$=$', '$\neg$', '$\cup$', '$\wedge$', '$\Box$', '$\Uparrow$', '$\Downarrow$', '$\Rightarrow$', '$\forall$', '(', ')', and the following, mutually disjoint, countable, sets of symbols:*

- *C_D, C_T, C_E (constants of sorts D, T and E),*
- *V_D, V_T, V_E, V_F, V_Φ (variables of each sort),*
- *F_D, F_T, F_E (function symbols of each arity $n \geq 1$ of sorts D, T, E), and*
- *R_D, R_E, R_F, R_Φ (relation symbols of each arity $n \geq 0$ of sorts D, E, F, and Φ).*

The constants of sorts F and Φ are defined below.

Definition 2. *The terms of each sort S are defined as follows:*

- *If S is of sort D or T then $term_S = C_S \cup V_S \cup \{f(u_1, \ldots, u_n) : n\text{-ary } f \in F_S, u_i \in term_S\}$.*
- *$term_E = C_E \cup V_E \cup \{f(u_1, \ldots, u_n) : n\text{-ary } f \in F_E, u_i \in term_D \cup term_T \cup term_E\}$.*
- *$term_F = C_F \cup V_F$, where $C_F = \{r_D(u_1, \ldots, u_n) : n\text{-ary } r_D \in R_D, u_i \in term_D\}$.*
- *$term_\Phi = C_\Phi \cup V_\Phi$, where $C_\Phi = \mathcal{EL}$ is defined in Definition 3.*

Definition 3. *$\mathcal{EL}$ is the minimal set which satisfies the following conditions.*

- *If $t, t' \in term_T$ then $t < t' \in \mathcal{EL}$.*
- *If S is of sort D, T, E or F, and $u, u' \in term_S$, then $u = u' \in \mathcal{EL}$.*

Table 1: Satisfaction and violation conditions for $\mathcal{CL}$ (see Definition 7)

$$
\begin{array}{rll}
M,w,g \models t < t' & \text{iff} & \langle \mathcal{V}_g(t), \mathcal{V}_g(t') \rangle \in \prec_{\mathcal{T}} \\
M,w,g \nmodels t < t' & \text{iff} & \langle \mathcal{V}_g(t), \mathcal{V}_g(t') \rangle \notin \prec_{\mathcal{T}} \\
M,w,g \models u = u' & \text{iff} & \mathcal{V}_g(u) \text{ is } \mathcal{V}_g(u') \\
M,w,g \nmodels u = u' & \text{iff} & \mathcal{V}_g(u) \text{ is not } \mathcal{V}_g(u') \\
M,w,g \models r_S(u_1,\ldots,u_n)(t) & \text{iff} & \mathcal{V}^R_S(r_S,w,\mathcal{V}_g(t))(\mathcal{V}_g(u_1),\ldots,\mathcal{V}_g(u_n)) = \textit{true} \\
M,w,g \nmodels r_S(u_1,\ldots,u_n)(t) & \text{iff} & \mathcal{V}^R_S(r_S,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 V_F \text{ and } M,w,g \models \mathcal{V}_g(v)(t) \\
M,w,g \nmodels v(t) & \text{iff} & v \in V_F \text{ and } M,w,g \nmodels \mathcal{V}_g(v)(t) \\
M,w,g \models r_\Phi(u_1,\ldots,u_n) & \text{iff} & \mathcal{R}_\Phi(r_\Phi,w)(\mathcal{V}_g(u_1),\ldots,\mathcal{V}_g(u_n)) = \textit{true} \\
M,w,g \nmodels r_\Phi(u_1,\ldots,u_n) & \text{iff} & \mathcal{R}_\Phi(r_\Phi,w)(\mathcal{V}_g(u_1),\ldots,\mathcal{V}_g(u_n)) = \textit{false} \\
M,w,g \models v & \text{iff} & v \in V_\Phi \text{ and } M,w,g \models \mathcal{V}_g(v) \\
M,w,g \nmodels v & \text{iff} & v \in V_\Phi \text{ and } M,w,g \nmodels \mathcal{V}_g(v) \\
M,w,g \models \neg\psi & \text{iff} & M,w,g \nmodels \psi \\
M,w,g \nmodels \neg\psi & \text{iff} & M,w,g \models \psi \\
M,w,g \models \mathsf{U}\psi & \text{iff} & \text{neither } M,w,g \models \psi \text{ nor } M,w,g \nmodels \psi \\
M,w,g \nmodels \mathsf{U}\psi & \text{iff} & \text{either } M,w,g \models \psi \text{ or } M,w,g \nmodels \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 \nmodels \psi \wedge \chi & \text{iff} & M,w,g \nmodels \psi \text{ or } M,w,g \nmodels \chi \\
M,w,g \models \Box\psi & \text{iff} & M,w',g \models \psi \text{ for every } w' \in \mathcal{W} \\
M,w,g \nmodels \Box\psi & \text{iff} & M,w',g \nmodels \psi \text{ for some } w' \in \mathcal{W} \\
M,w,g \models \psi \Uparrow \chi & \text{iff} & M,w,g \models \neg\mathsf{F}\psi \text{ and } \uparrow(\psi,w,g) \subseteq [\![\chi]\!]^M_g \\
M,w,g \nmodels \psi \Uparrow \chi & \text{iff} & M,w,g \models \neg\mathsf{F}\psi \text{ and } \uparrow(\psi,w,g) \bullet [\![\neg\mathsf{T}\chi]\!]^M_g \\
M,w,g \models \psi \Downarrow \chi & \text{iff} & M,w,g \models \mathsf{F}\psi \text{ and } \downarrow(\neg\mathsf{F}\psi,w,g) \subseteq [\![\psi \Uparrow \chi]\!]^M_g \\
M,w,g \nmodels \psi \Downarrow \chi & \text{iff} & M,w,g \models \mathsf{F}\psi \text{ and } \downarrow(\neg\mathsf{F}\psi,w,g) \bullet [\![\neg\mathsf{T}(\psi \Uparrow \chi)]\!]^M_g \\
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 \nmodels \psi \Rightarrow \chi & \text{iff} & M,w,g \nmodels \psi \Uparrow \chi \text{ or } M,w,g \nmodels \psi \Downarrow \chi \\
M,w,g \models \forall v\psi & \text{iff} & M,w,g' \models \psi \text{ for every } g' \text{ such that } g \approx_v g' \\
M,w,g \nmodels \forall v\psi & \text{iff} & M,w,g' \nmodels \psi \text{ for some } g' \text{ such that } g \approx_v g'
\end{array}
$$

- *If S is of sort D, E or F, $u_1,\ldots,u_n \in$ term$_S$, r_S is an n-ary relation symbol in R_S, and $t \in$ term$_T$, then $r_S(u_1,\ldots,u_n)(t) \in \mathcal{EL}$.*
- *If $v \in V_F$ and $t \in term_T$ then $v(t) \in \mathcal{EL}$.*
- *If $\phi,\psi \in \mathcal{EL}$, then $\neg\phi \in \mathcal{EL}$, $\mathsf{U}\phi \in \mathcal{EL}$, and $(\phi \wedge \psi) \in \mathcal{EL}$.*
- *If S is of sort D, T, E or F, $v \in V_S$ and $\phi \in \mathcal{EL}$, then $\forall v\phi \in \mathcal{EL}$.*

The members of $\mathcal{EL}$ are called formulas *(of $\mathcal{EL}$). Those formulas in which no variable occurs free are called* sentences *(of $\mathcal{EL}$).*

Definition 4. *$\mathcal{CL}$ is the minimal set which satisfies the following conditions.*

- *$\mathcal{EL} \subseteq \mathcal{CL}$.*
- *If $u,u' \in$ term$_\Phi$, then $u = u' \in \mathcal{CL}$.*
- *If $u_1,\ldots,u_n \in$ term$_\Phi$ and r_Φ is an n-ary relation symbol in R_Φ, then $r_\Phi(u_1,\ldots,u_n) \in \mathcal{CL}$.*
- *If $\phi,\psi \in \mathcal{CL}$, then $\neg\phi \in \mathcal{CL}$, $\mathsf{U}\phi \in \mathcal{CL}$, $(\phi \wedge \psi) \in \mathcal{CL}$, $\Box\phi \in \mathcal{CL}$, $(\phi \Uparrow \psi) \in \mathcal{CL}$, $(\phi \Downarrow \psi) \in \mathcal{CL}$, and $(\phi \Rightarrow \psi) \in \mathcal{CL}$.*
- *If S is any sort, $v \in V_S$ and $\phi \in \mathcal{CL}$, then $\forall v\phi \in \mathcal{CL}$.*

The members of $\mathcal{CL}$ are called formulas *(of $\mathcal{CL}$). Those formulas in which no variable occurs free are called* sentences *(of $\mathcal{CL}$).*

Models of $\mathcal{CL}$ consist of a possible partial worlds frame $\langle \mathcal{W}, \prec \rangle$, a set $\mathcal{D}$ of domain objects, a set $\mathcal{E}$ of event types, a temporal frame $\langle \mathcal{T}, \prec_{\mathcal{T}} \rangle$ (where $\mathcal{T}$ is a set of time points and $\prec_{\mathcal{T}}$ is the before-after relation on $\mathcal{T}$), and interpretation functions for terms and relations. For simplicity, time is assumed to be discrete and linear. The denotations of terms are always defined and do not vary with time. By contrast

relations are interpreted by time-dependent, partial, characteristic functions; thus the interpretation of relations may be partial and may vary with time.

Definition 5. *A* model *for $\mathcal{CL}$ is a structure*

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

where

- *$\mathcal{W}$ is a set and $\prec$ is a partial order on $\mathcal{W}$,*
- *$\mathcal{D}$, $\mathcal{E}$ and $\mathcal{T}$ are mutually disjoint, countable, non-empty sets,*
- *$\prec_{\mathcal{T}}$ 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, for each pair $\langle S, \mathcal{S}\rangle \in \{\langle D, \mathcal{D}\rangle, \langle T, \mathcal{T}\rangle, \langle E, \mathcal{E}\rangle\}$, $\mathcal{F}_S$ is a set of n-ary functions of type $\mathcal{S}^n \rightarrow \mathcal{S}$ for each $n \geq 1$,*
- *$\mathcal{R} = \langle \mathcal{R}_D, \mathcal{R}_E, \mathcal{R}_F, \mathcal{R}_\Phi\rangle$, where for each pair $\langle S, \mathcal{S}\rangle \in \{\langle D, \mathcal{D}\rangle, \langle E, \mathcal{E}\rangle, \langle F, C_F\rangle, \langle \Phi, C_\Phi\rangle\}$, $\mathcal{R}_S$ is a set of partial n-ary functions of type $\mathcal{S}^n \rightarrow \{true, false\}$ for each $n \geq 0$,*
- *$\mathcal{V} = \langle\langle \mathcal{V}_D^C, \mathcal{V}_T^C, \mathcal{V}_E^C, \mathcal{V}_F^C, \mathcal{V}_\Phi^C\rangle, \langle \mathcal{V}_D^F, \mathcal{V}_T^F, \mathcal{V}_E^F\rangle, \langle \mathcal{V}_D^R, \mathcal{V}_E^R, \mathcal{V}_F^R\rangle\rangle$ is an interpretation function such that*
  - *$\mathcal{V}_S^C : C_S \rightarrow \mathcal{S}$ and $\mathcal{V}_S^F : F_S \rightarrow \mathcal{F}_S$ for $\langle S, \mathcal{S}\rangle \in \{\langle D, \mathcal{D}\rangle, \langle T, \mathcal{T}\rangle, \langle E, \mathcal{E}\rangle\}$,*
  - *$\mathcal{V}_F^C : C_F \rightarrow C_F$ and $\mathcal{V}_\Phi^C : C_\Phi \rightarrow C_\Phi$ are identity functions,*
  - *$\mathcal{V}_S^R : R_S \times \mathcal{T} \rightarrow \mathcal{R}_S$.*

Terms are interpreted in the standard way.

Definition 6. *A* variable assignment *for a $\mathcal{CL}$-model is a function $g = \langle g_D, g_T, g_E, g_F\rangle$, where for $\langle S, \mathcal{S}\rangle \in \{\langle D, \mathcal{D}\rangle, \langle T, \mathcal{T}\rangle, \langle E, \mathcal{E}\rangle, \langle F, C_F\rangle, \langle \Phi, C_\Phi\rangle\}$, $g_S : V_S \rightarrow \mathcal{S}$. For $\mathcal{CL}$-model M, interpretation function $\mathcal{V}$ and variable assignment g for M, the* term evaluation function *$\mathcal{V}_g$ is defined, for each $\mathcal{CL}$-term u and sort S, as follows*

$$\mathcal{V}_g(u) = \begin{cases} \mathcal{V}_S(u) & \text{if } u \in C_S, \\ g_S(u) & \text{if } u \in V_S, \\ \mathcal{V}_S^F(f)(\mathcal{V}_g(u_1), \ldots, \mathcal{V}_g(u_n)) & \text{otherwise.} \end{cases}$$

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 7. *Let $M = \langle\langle \mathcal{W}, \prec\rangle, \mathcal{D}, \mathcal{E}, \langle \mathcal{T}, \prec_{\mathcal{T}}\rangle, \mathcal{F}, \mathcal{R}, \mathcal{V}\rangle$ be a $\mathcal{CL}$-model, g be a variable assignment for M, and ϕ be a $\mathcal{CL}$-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 $M, w, g \models\!\!\!| \phi$) according to the clauses given in Table 1. In the table a number of abbreviations are used: $g \approx_v g'$ indicates that the variable assignments g and g' differ at most on the assignment to variable v; $[\![\phi]\!]_g^M$ denotes the set $\{w : M, w, g \models \phi\}$ of all worlds in M at which g satisfies ϕ; for $\mathcal{CL}$-formula ϕ, world w and assignment g, $\uparrow(\phi, w, g)$ denotes the set $\{w' : w \preceq w' \wedge w' \in [\![\phi]\!]_g^M \wedge \neg\exists w''(w \preceq w'' \prec w' \wedge w'' \in [\![\phi]\!]_g^M)\}$ of closest worlds above w at which g satisfies ϕ, and $\downarrow(\phi, w, g)$ denotes the set $\{w' : w' \preceq w \wedge w' \in [\![\phi]\!]_g^M \wedge \neg\exists w''(w' \prec w'' \preceq w \wedge w'' \in [\![\phi]\!]_g^M)\}$ of closest worlds below w at which g satisfies ϕ; finally, for sets S and T, $S \bullet T$ ("S overlaps T") iff $S \cap T \neq \emptyset$.*

A formula ϕ is true *at a possible partial world w in a $\mathcal{CL}$-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 $M, w \models\!\!\!| \phi$) if $M, w, g \models\!\!\!| \phi$ for all variable assignments g. A formula ϕ is* true *in a $\mathcal{CL}$-model M (written $M \models \phi$) if ϕ is true at all worlds in M. Formula ϕ is* false *in M (written $M \models\!\!\!| \phi$) if ϕ is false at all worlds in M. These definitions are extended to sets of formulas in the normal way. Thus the set of formulas Θ is* true *at a possible partial world w in a $\mathcal{CL}$-model M (written $M, w \models \Theta$) if $M, w \models \phi$ for all $\phi \in \Theta$, etc.*

It can be proved (by means of a parallel induction on the structure of $\mathcal{CL}$ formulas) that for any $\mathcal{CL}$-model M, world w in M, and $\mathcal{CL}$-sentence ϕ: either $M, w \models \phi$ or $M, w \models \neg\phi$ or $M, w \models \mathsf{U}\phi$. Consequently, as in classical logic, it is sufficient to consider the truth relations on sentences (and sets of sentences) of $\mathcal{CL}$.

Events

The theory of events begins with *primary events*, which can be thought of as defeasible STRIPS events (Fikes & Nilsson 1971). Primary event types are defined by specifying their preconditions and effects; examples are given in Section 6. The preconditions can be thought of as necessary conditions for the success of an event of this type, and the effects as its invariant effects. Consequently, the definition of an event (of type) e is said to be *natural* if its preconditions do not succeed its occurrence (if the definition of *Pre*$(e)(t)$ does not imply any literal $\ell(t')$ for any $t' > t$) and its effects do (if the definition of *Eff*$(e)(t)$ does not imply any literal $\ell(t')$ for any $t' < t$). The preconditions of an event should normally be sufficient, on its occurrence, for its effects, but will typically not logically guarantee them. Call the context in which an event occurs its *context of occurrence*. Then the preconditions of the event should be such that they are sufficient in most contexts of occurrence, but need not be sufficient in all of them. In order to represent this, *success atoms* are introduced. Intuitively the *success atom*, *Succ*$(e)(t)$, states that event e succeeds at time t; that is, that e occurs at t, its preconditions are true on occurrence, and its effects are true at $t + 1$. This is stated by the success axiom, Axiom (1) in Table 2.

This axiom is intended to be used in order to infer change. Given *Occ*$(e)(t)$ and *Pre*$(e)(t)$, the *success assumption*, *Succ*$(e)(t)$, should be made whenever it is consistent to do so (whenever it is consistent with e's context of occurrence), and the axiom used to conclude *Eff*$(e)(t + 1)$.

Primary events have the defeasibility of natural events, but are unlike natural events in that their effects, when successful, are invariant. This limitation is overcome by introducing *secondary events*. Secondary events are defeasible STRIPS events which are *invoked* by other (primary or secondary) events in appropriate contexts, and their success depends on that of the events which invoke them. A common sense event can thus be thought of as a tree-structured object whose root is a primary event, and whose effects are the combined effects of all successful events in its invocation tree. Invocations are represented by *invocation*

Table 2: The theory of events, Θ_E

$$\forall e, t(\mathit{Succ}(e)(t) \equiv \mathsf{T}(\mathit{Occ}(e)(t) \wedge \mathit{Pre}(e)(t) \wedge \mathit{Eff}(e)(t+1))) \quad (1)$$
$$\forall e, e', t(\mathit{Inv}(e, e')(t) \rightarrow \neg \mathit{Inv}(e', e)(t)) \quad (2)$$
$$\forall e, e', t(\mathit{Inv}(e, e')(t) \rightarrow (\mathit{Occ}(e)(t) \wedge \mathit{Occ}(e')(t))) \quad (3)$$
$$\forall e, e', t((\mathit{Inv}(e, e')(t) \wedge \mathit{Succ}(e')(t)) \rightarrow \exists e''(\mathit{Inv}(e'', e')(t) \wedge \mathit{Succ}(e'')(t))) \quad (4)$$
$$\forall e, e', e'', t(\mathit{Inv}^*(e, e')(t) \equiv (\mathit{Inv}(e, e')(t) \vee (\mathit{Inv}(e, e'')(t) \wedge \mathit{Inv}^*(e'', e')(t)))) \quad (5)$$
$$\forall \alpha, t(\mathit{Inert}(\alpha)(t) \equiv (\alpha(t) \equiv \alpha(t+1))) \quad (6)$$

Table 3: The theory of causation, $\Theta_C(\theta)$

$$\forall \epsilon, e, t, \phi, \psi(\mathit{PSCause}(\epsilon, \phi) \equiv (\epsilon = \mathit{Occ}(e)(t) \wedge \mathit{Succ}(e)(t) \wedge \Box(\theta \rightarrow (\mathit{Eff}(e)(t+1) \equiv \psi)) \wedge \Box(\psi \rightarrow \phi))) \quad (7)$$
$$\forall \epsilon, \phi, e, e', t(\mathit{CSCause}(\epsilon, \phi) \equiv (\epsilon = \mathit{Occ}(e)(t) \wedge \phi = \mathit{Occ}(e')(t) \wedge \mathit{Succ}(e)(t) \wedge \mathsf{T}\mathit{Inv}(e, e')(t) \wedge \neg \mathsf{T} \exists e''(\mathit{Inv}^*(e'', e)(t) \wedge \mathit{Inv}(e'', e')(t)))) \quad (8)$$
$$\forall \epsilon, \phi, \psi, \chi(\mathit{SCause}(\epsilon, \phi) \equiv (\mathit{PSCause}(\epsilon, \phi) \vee \mathit{CSCause}(\epsilon, \phi) \vee (\phi = (\psi \wedge \chi) \wedge \mathit{SCause}(\epsilon, \psi) \wedge \mathit{SCause}(\epsilon, \chi)))) \quad (9)$$
$$\forall \epsilon, \phi(\mathit{Cause}(\epsilon, \phi) \equiv (\mathit{SCause}(\epsilon, \phi) \wedge \neg \exists \epsilon'(\neg \epsilon' = \epsilon \wedge \mathit{SCause}(\epsilon', \phi)) \wedge (\neg \mathsf{T}\epsilon \Downarrow (\neg \mathsf{T}\phi \vee \exists \epsilon' \mathit{Cause}(\epsilon', \phi))))) \quad (10)$$
$$\forall \epsilon, \epsilon', \phi, \psi, \chi(\mathit{Causes}(\epsilon, \phi) \equiv (\mathit{Cause}(\epsilon, \phi) \vee (\mathit{Cause}(\epsilon, \epsilon') \wedge \mathit{Causes}(\epsilon', \phi)) \vee (\phi = (\psi \wedge \chi) \wedge \mathit{Causes}(\epsilon, \psi) \wedge \mathit{Causes}(\epsilon, \chi)))) \quad (11)$$

atoms, thus the atom $\mathit{Inv}(e, e')(t)$ states that event e invokes event e' at time t, and by invocation axioms of the form: $\forall e, e', t((\mathit{Occ}(e)(t) \wedge \Phi(e, e')(t)) \rightarrow \mathit{Inv}(e, e')(t))$; where $\Phi(e, e')(t)$ is a formula which distinguishes those contexts in which e invokes e' at t. In keeping with the suggested properties of secondary events, the invocation relation is required to satisfy axioms (2)-(4) in Table 2. In particular, Axiom (4) ensures that a secondary event succeeds only if it is directly invoked by a successful event.

It is also necessary to represent inertia, or what is not changed by events. Intuitively, the *inertia atom*, $\mathit{Inert}(\alpha)(t)$, states that the truth value of fact α does not change at time t; that is, that it persists to $t+1$. This is stated by the inertia axiom, Axiom (6). The intention is that the inertia axiom should be used to infer persistence of facts whenever possible. Given $\alpha(t)$, the *inertia assumption*, $\mathit{Inert}(\alpha)(t)$, should be made whenever it is consistent to do so (given the context of occurrence at t), and the axiom used to conclude $\alpha(t+1)$.

Definition 8. *The* theory of events, Θ_E, *consists of the axioms in Table 2; thus* $\Theta_E = \{(1), \ldots, (6)\}$. *An* event theory *is any set of $\mathcal{EL}$ sentences which contains Θ_E. An event theory is* natural *if every event definition it contains is natural.*

This theory is capable of representing inertia, ramifications, qualifications, and non-determinism (Bell 2000; 2003), and can readily be extended in order to represent conflicting simultaneous events (Bell 2001b).

Causation

Causation is defined formally in Table 3 relative to an $\mathcal{EL}$-sentence θ, which represents a set of background laws; including the definitions of the preconditions and effects of events, and domain constraints. The definition assumes the background of the theory of events, and reduces the notion of causation to its terms (event occurrences, preconditions, effects, success, failure, invocations, facts, inertia, change), together with the modal notions of physical and contextual necessity.

Axiom (7) states that any event e which succeeds at time t is a *prior sufficient cause* (a *PSCause*) of its (direct posterior) effects. Thus e is a *PSCause* of ϕ if e succeeds at t and ϕ is a physically necessary consequence of e's effects at $t+1$ according to the background laws θ.

Axiom (8) states that the occurrence of event e is a *contemporaneous sufficient cause* (a *CSCause*) of the occurrence of event e' at t iff e succeeds and invokes e' at t, and it is not true that there is an event e'' which (directly or indirectly) invokes e and which also (directly) invokes e' at t. This requirement ensures that contemporaneous sufficient causation is attributed to the earliest invoking event in an invocation chain.

More abstractly, Axiom (9) states that event occurrence ϵ is a (contextually) *sufficient cause* (an *SCause*) of effect ϕ iff ϵ is a prior sufficient cause of ϕ, or a contemporaneous sufficient cause of ϕ, or a sufficient cause of both ψ and χ whose conjunction is ϕ.

The occurrence ϵ is a *direct cause* (a *Cause*) of effect ϕ iff ϵ is the only sufficient cause of effect ϕ in the context, and removing ϵ from the context either removes ϕ or results in some other (trumped, premempted) occurrence causing ϕ; Axiom (10).

Indirect causation results from causally linked chains of events, each of which may terminate in a fact. Accordingly, the indirect-causation relation *Causes* is defined to be the transitive closure of the direct-causation relation *Cause*, and is closed under conjunction of effects; Axiom (11).

Definition 9. *The* theory of causation *relative to $\mathcal{EL}$-sentence θ, $\Theta_C(\theta)$, consists of θ together with the axioms given in Table 3; thus* $\Theta_C(\theta) = \{\theta, (7), \ldots, (11)\}$. *If* $\Theta' = \Theta_E \cup \Theta_L \cup \Theta_B$ *is a finite event theory with background laws Θ_L and boundary conditions Θ_B, then* $\Theta = \Theta' \cup \Theta_C(\bigwedge \Theta_L)$

Table 4: Properties of the *Causes* relation (see Proposition 1)

Bivalence:	$\forall \epsilon, \phi(\mathit{Causes}(\epsilon, \phi) \vee \neg \mathit{Causes}(\epsilon, \phi))$
Transitivity:	$\forall \epsilon, \epsilon', \phi((\mathit{Causes}(\epsilon, \epsilon') \wedge \mathit{Causes}(\epsilon', \phi)) \rightarrow \mathit{Causes}(\epsilon, \phi))$
Asymmetry:	$\forall \epsilon, \phi(\mathit{Causes}(\epsilon, \phi) \rightarrow \neg \mathit{Causes}(\phi, \epsilon))$
Actuality:	$\forall \epsilon, \phi(\mathit{Causes}(\epsilon, \phi) \rightarrow (\epsilon \wedge \phi))$
Consistency:	$\forall \epsilon, \phi(\mathit{Causes}(\epsilon, \phi) \rightarrow \neg \mathit{Causes}(\epsilon, \neg \mathsf{T}\phi))$
Conjunction:	$\forall \epsilon, \phi, \psi((\mathit{Causes}(\epsilon, \phi) \wedge \mathit{Causes}(\epsilon, \psi)) \rightarrow \mathit{Causes}(\epsilon, \phi \wedge \psi))$

is a causal theory, *and* $\mathit{Laws}(\Theta) = \{\Theta_E \cup \Theta_L \cup \Theta_C(\bigwedge \Theta_L)\}$ *is the set of* laws of Θ.

The definition of causation has appropriate general properties.

Proposition 1. *Let Θ be a consistent causal theory whose constituent event theory is natural. Then each of the sentences listed in Table 4 is true in every model of Θ.*

Proof Sketch. The case for Conjunction follows directly from the definition; Axiom (11). For Transitivity, suppose $\mathit{Causes}(\epsilon, \epsilon')$, $\mathit{Causes}(\epsilon', \phi)$ and $\neg\epsilon = \epsilon'$. If $\mathit{Cause}(\epsilon, \epsilon')$, then the conclusion follows by Axiom (11). Otherwise, there is some ϵ'' such that $\mathit{Cause}(\epsilon, \epsilon'')$ and $\mathit{Causes}(\epsilon'', \epsilon')$. Supposing $\mathit{Causes}(\epsilon'', \phi)$ as induction hypothesis, we conclude by Axiom (11). Actuality follows from the definition of *SCause*, for if $\mathit{SCause}(\epsilon, \phi)$ is true, then so are both ϵ and ϕ (axioms (1), (3), (7), (8) (9)). Consistency follows from Actuality and the assumption that Θ is consistent. For Bivalence, note first that *SCause* is bivalent. If $\mathit{SCause}(\epsilon, \phi)$ is false, then so is $\mathit{Cause}(\epsilon, \phi)$; Axiom (10). Otherwise, ϵ is actual and so $\mathsf{F}\neg\mathsf{T}\epsilon$ is true. The contrafactual of Axiom (10) is thus applicable and is consequently either true or false. Finally, Asymmetry follows because the naturalness assumption and axioms (2) and (9) ensure that *SCause* is asymmetric. □

Causal Models

We begin by defining the set of preferred worlds for an event theory in a $\mathcal{CL}$-model M; intuitively these are the worlds in M at which the event theory is interpreted as intended, and events unfold according to the theory. The definition refines Shoham's (1988) idea of chronological minimization. The preferred worlds of an event theory are those in which defined atomic sentences are minimized chronologically with type-defined priority at each time point. The effect is that at each preferred world the event theory is interpreted chronologically. At each time point in a preferred world the context of occurrence is first fixed (by minimizing the facts and event structures which are defined at that point), then the events occurring in the context are assumed to succeed wherever possible (success assumptions are maximized), and finally the facts in the context are assumed to persist whenever possible (inertia assumptions are maximized); further motivation and justification is given in (Bell 2000).

Definition 10. *Let M be a $\mathcal{CL}$ model with world set $\mathcal{W}$. Then the* event preference relation for M, $\prec_E^M$, *is defined as follows. For all $w, w' \in \mathcal{W}$, put $w \prec_E^M w'$ iff there is a time point t such that w and w' agree on the interpretation of all atomic sentences at any earlier time point and either*

1. *at least one more domain atom or occurs atom is defined (is either true or false) at w' at t, or*
2. *w and w' agree on the interpretation of domain and occurs atoms at t, and at least one more invocation atom is defined at w' at t, or*
3. *w and w' agree on the interpretation of domain, occurs and invocation atoms at t, and at least one more success atom is false at w' at t, or*
4. *w and w' agree on the interpretation of domain, occurs, invocation and success atoms at t, and at least one more inertia atom is false at w' at t, or*
5. *w and w' agree on the interpretation of domain, occurs, invocation, success, and inertia atoms at t, and at least one more $\mathcal{EL}$-formula atom or event atom is defined at w' at t.*

*World w is an E-*preferred world *for a sentence ϕ in M iff $M, w \models \phi$ and there is no other world w' such that $M, w' \models \phi$ and $w' \prec_E^M w$. Similarly w is an E-preferred world for a set of sentences Θ in M iff $M, w \models \Theta$ and there is no other world w' such that $M, w' \models \Theta$ and $w' \prec_E^M w$.*

We turn now to the definition of the *causal model* for a causal theory. As we are not concerned with the interpretation of terms, we can simply stipulate that they are interpreted autonymously (as themselves). This fixes their meaning an allows us to consider a single model of the causal theory.

Intitively, the causal model for causal theory $\Theta = \mathit{Laws}(\Theta) \cup \Theta_B$ should contain all $\mathit{Laws}(\Theta)$-worlds at which the laws are interpreted as intended and at which the boundary conditions may vary, and these worlds should be ordered according to their similarity to the actual world. Consequently the model should contain the set of E-preferred worlds for each causal theory $\Theta' = \mathit{Laws}(\Theta) \cup \Theta'_B$. Moreover, if w is an E-preferred Θ'-world, then the closest ϕ-worlds above w in the model should be those at which w's boundary conditions Θ'_B are changed as little as possible in order that ϕ is true, and similarly for the closest $\neg\mathsf{T}\phi$-worlds below w. However, closeness in this sense cannot simply be defined in terms of the $\prec_E^M$ order, because neighbouring worlds may change in arbitrary ways, reflecting arbitrary changes in boundary conditions. For example, suppose that event e invokes event e' at world w and that e is the only event which does so. Then, as the occurrence of e' depends on the occurrence of e, the closest worlds below w at which e does not occur should also be worlds at which e' does not occur. However the $\prec_E^M$-closest worlds below w at which e does not occur are worlds at which e' does occur. In effect, the occurrence of e' is an additional boundary condition at

these worlds. This consideration leads to the idea of a *support set* for a world. Intuitively, a support set for world w is a minimal set of ground $\mathcal{EL}$-literals which support the boundary conditions at w.

Definition 11. *A ground $\mathcal{EL}$-literal is an atomic $\mathcal{EL}$-sentence $\alpha(t)$ or the (strong) negation $\neg\alpha(t)$ of an atomic $\mathcal{EL}$-sentence. Let $\mathcal{B}$ be the set of all subsets of ground $\mathcal{EL}$-literals, M be a $\mathcal{CL}$-model with event preference relation $\prec_E^M$, and let w be a world in M. Then set $B \in \mathcal{B}$* supports *w iff w is an E-preferred B-world in M and there is no set $B' \subset B$ such that w is an E-preferred B'-world in M.*

The subset relation on support sets gives the closeness relation we desire.

Definition 12. *Let $\Theta = \mathit{Laws}(\Theta) \cup \Theta_B$ be a causal theory. The* causal model *for Θ is the model in which terms are interpreted autonymously, with world set*

$$\mathcal{W} = \{w : B \in \mathcal{B} \text{ and } w \text{ is a } \mathit{Laws}(\Theta) \cup B\text{-world}\},$$

where $\mathcal{B}$ is as in Definition 11, and with accessibility relation $\prec_C^\Theta$ on $\mathcal{W}$ defined as follows

$$w \prec_C^\Theta w' \text{ iff there exist } B, B' \in \mathcal{B} \text{ such that } B \subset B', \\ B \text{ supports } w, \text{ and } B' \text{ supports } w'.$$

Clearly any causal theory Θ has a unique causal model M. Note that the world set of M may be empty if *Laws*(Θ) is inconsistent. Note also that if any *Laws*$(\Theta) \cup B$ is inconsistent, then M contains no *Laws*$(\Theta) \cup B$-world. Moreover, as all worlds in M are *Laws*(Θ)-worlds, it follows that all E-preferred B-worlds in M are also E-preferred *Laws*$(\Theta) \cup B$-worlds.

The following proposition follows straightforwardly from the definitions, and relates the evaluation of conditionals in causal models to the AGM expansion and contraction operations. Thus if $\neg\mathsf{F}\phi$ is true at w, then, a closest ϕ-world w' above w is obtained by adding a minimal set of literals to the support set of w such that ϕ is true. And if $\mathsf{F}\phi$ is true at w, then, a closest ϕ-world w' below w is obtained by removing a minimal set of literals from the support set of w such that $\neg\mathsf{F}\phi$ is true.

Proposition 2. *Let Θ be a causal theory with causal model M with event preference relation $\prec_E^M$ and accessibility relation $\prec_C^\Theta$. Then, for world w in M and $\mathcal{EL}$-sentence ϕ,*

1. *if $\neg\mathsf{F}\phi$ is true at w, then world w' is a $\prec_C^\Theta$-closest ϕ-world above w iff w has support set B, w' is a ϕ-world with support set B' such that $B \subseteq B'$, and there is no ϕ-world w'' with support set B'' such that $B \subseteq B'' \subset B'$.*
2. *if $\mathsf{F}\phi$ is true at w, then world w' is a $\prec_C^\Theta$-closest $\neg\mathsf{F}\phi$-world below w iff w has support set B, w' is a $\neg\mathsf{F}\phi$-world with support set B' such that $B' \subset B$, and there is no $\neg\mathsf{F}\phi$-world w'' with support set B'' such that $B' \subset B'' \subset B$.*

Causal entailment can now be defined as follows.

Definition 13. *Let Θ be a causal theory with causal model M, event preference relation $\prec_E^M$. Then Θ is* pragmatically consistent *if M contains at least one E-preferred Θ-world, Θ is* determinisitic *if M contains exactly one E-preferred Θ-world, Θ is* nondeterminisitic *if M contains more than one E-preferred Θ-world, and Θ* causally entails *a sentence ϕ, written $\Theta \mid\!\!\approx_C \phi$, iff every E-preferred Θ-world in M is also a ϕ-world.*

Thus counterfactual reasoning in a causal model consists of using its event-preference relation to find the actual world (or, if the theory is nondeterministic, the set of worlds which represent the actual world), and then using the accessiblity relation to find closest antecedent worlds.

Examples

In order to illustrate the workings of the theory, and particularly the contextual necessity condition, examples are now given of preemption and trumping preemption.

Example 1. Wright's (1988) two-fires example can be stated more specifically as follows. Fires A and B break out at opposite ends of a terrace of four houses, $H_1, \ldots, H_4$. A starts in H_1 and spreads to H_2. Simultaneously, B starts in H_4 and spreads to H_3. Consequently A destroys H_2 before B can reach it. In the circumstances it seems natural to conclude that A is the cause of H_2's destruction, depite the fact that, if A had not occurred, then B would have spread from H_3 to H_2 and caused H_2's destruction at a later time.

This version of the example can be represented by adding the axioms (12)-(18) from Table 5. Axioms (12)-(15) state (simplified) preconditions and effects for *Fire* and *Spreads* events. In particular the preconditions for a fire spreading from location l to location l' are that the fire is currently burning at l, that l and l' are adjacent, that l' has not already been burnt (that is, that whatever is at l is combustible), and that there is not currently a fire burning at l'. Axiom (16) is an invocation axiom which states that a *Fire* event invokes a *Spreads* event if the preconditions of the *Spreads* event are true in the context in which the *Fire* event occurs. Finally, axioms (17) and (18) state the boundary conditions. In Axiom (17), the notation *UN*$[u_1, \ldots, u_n]$ indicates that each of the names $u_1, \ldots, u_n$ refers to a distinct individual: *UN*$[u_1, \ldots, u_n] =_{\mathrm{Df}} \bigwedge \neg u_i = u_j$ for $1 \leq i < j \leq n$.

Let Θ be the causal theory with background laws $\Theta_L = \{(12), \ldots, (16)\}$ and boundary conditions $\Theta_B = \{(17), (18)\}$. Then

$$\Theta \mid\!\!\approx_C \mathit{Causes}(\mathit{Occ}(\mathit{Fire}(A, H_1))(1), \mathit{Burnt}(H_2)(3)) \\ \wedge (\neg\mathsf{T}\mathit{Occ}(\mathit{Fire}(A, H_1))(1) \Downarrow \\ \mathit{Causes}(\mathit{Occ}(\mathit{Fire}(B, H_4))(1), \mathit{Burnt}(H_2)(4))).$$

Proof. Let M be the causal model for Θ with event preference relation $\prec_E^M$ and accessibility relation $\prec_C^\Theta$. There is a single E-preferred Θ-world, w, in M; Θ is thus deterministic. By axioms (12) and (18), *Occ*(*Fire*$(A, H_1))(1)$ and *Pre*(*Fire*$(A, H_1))(1)$ are true (at w). It follows from axioms (14), (17), (18), and the minimization of occurrences at time 1 that *Pre*(*Spreads*$(A, H_1, H_2))(1)$ is true. So it follows by axioms (3) and (16) that both *Inv*(*Fire*(A, H_1), *Spreads*$(A, H_1, H_2))(1)$ and *Occ*(*Spreads*$(A, H_1, H_2))(1)$ are true. It follows from the maximization of successes at time 1 that *Succ*(*Fire*$(A, H_1))(1)$ is true. And so it follows from axioms (8) and (9) that *SCause*(*Occ*(*Fire*$(A, H_1))(1)$, *Occ*(*Spreads*$(A, H_1, H_2))(1))$ is true. Moreover, it follows

Table 5: Axioms for the examples

$\forall x, l, t(\mathit{Pre}(\mathit{Fire}(x, l))(t) \equiv \neg \mathit{Burnt}(l)(t))$ (12)

$\forall x, l, t(\mathit{Eff}(\mathit{Fire}(x, l))(t) \equiv \mathit{Burnt}(l)(t))$ (13)

$\forall x, l, l', t(\mathit{Pre}(\mathit{Spreads}(x, l, l'))(t) \equiv$
$(\mathit{Occ}(\mathit{Fire}(x, l))(t) \wedge \mathit{Adj}(l, l')(t) \wedge \mathit{Pre}(\mathit{Fire}(x, l'))(t) \wedge \neg \mathsf{T} \exists y\, \mathit{Occ}(\mathit{Fire}(y, l'))(t)))$ (14)

$\forall x, l, l', t(\mathit{Eff}(\mathit{Spreads}(x, l, l'))(t) \equiv \mathit{Occ}(\mathit{Fire}(x, l'))(t))$ (15)

$\forall x, l, l', t((\mathit{Occ}(\mathit{Fire}(x, l))(t) \wedge \mathit{Pre}(\mathit{Spreads}(x, l, l'))(t)) \rightarrow \mathit{Inv}(\mathit{Fire}(x, l), \mathit{Spreads}(x, l, l'))(t))$ (16)

$\mathit{UN}[H_1, \ldots, H_4] \wedge \bigwedge_{i=1}^{3} (\mathit{Adj}(H_i, H_{i+1})(1) \wedge \mathit{Adj}(H_{i+1}, H_i))(1)) \wedge \bigwedge_{i=1}^{4} \neg \mathit{Burnt}(H_i)(1)$ (17)

$\mathit{Occ}(\mathit{Fire}(A, H_1))(1) \wedge \mathit{Occ}(\mathit{Fire}(B, H_4))(1)$ (18)

$\forall x, y, e, t(\mathit{Pre}(\mathit{Ord}(x, y, e))(t) \equiv (\mathit{OutRanks}(x, y)(t) \wedge \forall z, e'((\neg z = x \wedge \mathit{Ord}(z, y, e')(t)) \rightarrow \mathit{OutRanks}(x, z)(t))))$ (19)

$\forall x, y, e, t(\mathit{Eff}(\mathit{Ord}(x, y, e))(t) \equiv \mathit{Occ}(e)(t))$ (20)

$\mathit{UN}[S_1, S_2, S_3] \wedge \mathit{OutRanks}(S_1, S_2)(1) \wedge \mathit{OutRanks}(S_1, S_3)(1) \wedge \mathit{OutRanks}(S_2, S_3)(1)$ (21)

$\mathit{Occ}(\mathit{Ord}(S_1, S_3, \mathit{Adv}(S_3)))(1) \wedge \mathit{Occ}(\mathit{Ord}(S_2, S_3, \mathit{Adv}(S_3)))(1)$ (22)

from the definition of *SCause* and the minimization of occurrences at time 1, that no other occurrence ϵ is an *SCause* of A spreading to H_1 at time 1; thus $\neg\exists\epsilon(\neg\epsilon = \mathit{Occ}(\mathit{Fire}(A, H_1))(1) \wedge \mathit{SCause}(\epsilon, \mathit{Occ}(\mathit{Spreads}(A, H_1, H_2))(1)))$ is true at w.

The support set for w, B, consists of the ground literals which support axioms (17) and (18). In particular, the event literals in B consist of the two *Fire* events in the latter axiom. So, in view of Proposition 2, w', the $\prec_C^{\Theta}$-closest $\neg\mathsf{T}\mathit{Occ}(\mathit{Fire}(A, H_1))(1)$-world below w is obtained by removing $\mathit{Occ}(\mathit{Fire}(A, H_1))(1)$ from B. Doing so leaves $\mathit{Inv}(\mathit{Fire}(A, H_1), \mathit{Spreads}(A, H_1, H_2))(1)$ and $\mathit{Occ}(\mathit{Spreads}(B, H_4, H_3))(1)$ unsupported, and hence, by minimization of occurrences, $\neg\mathsf{T}\mathit{Occ}(\mathit{Spreads}(B, H_1, H_2))(1)$ is true at w'. It follows that $\neg\mathsf{T}\mathit{Occ}(\mathit{Fire}(A, H_1))(1) \Downarrow \neg\mathsf{T}\mathit{Occ}(\mathit{Spreads}(A, H_1, H_2))(1)$ is true at w. So, by Axiom (10), $\mathit{Cause}(\mathit{Occ}(\mathit{Fire}(A, H_1))(1), \mathit{Occ}(\mathit{Spreads}(A, H_1, H_2))(1))$ is true at w.

Similar reasoning shows that $\mathit{Occ}(\mathit{Fire}(A, H_4))(1)$ is the *Cause* of $\mathit{Occ}(\mathit{Spreads}(B, H_4, H_3))(1)$ at w.

It follows from the success of $\mathit{Occ}(\mathit{Fire}(A, H_1))(1)$ and maximization of successes that $\mathit{Occ}(\mathit{Spreads}(A, H_1, H_2))(1)$ succeeds at w at time 1, and so by axioms (1) and (15), that $\mathit{Occ}(\mathit{Fire}(A, H_2))(2)$ is true at w. By axioms (7) and (9), $\mathit{Occ}(\mathit{Spreads}(A, H_1, H_2))(1)$ is an *SCause* of $\mathit{Occ}(\mathit{Fire}(A, H_2))(2))$ at w, and by the above uniqueness argument, it is the only *SCause*.

Now the $\prec_C^{\Theta}$-closest $\neg\mathsf{T}\mathit{Occ}(\mathit{Spreads}(A, H_1, H_2))(1))$-world below w is the world w' referred to above. Removing the occurrence of the $\mathit{Fire}(A, H_1)$-event from B leaves $\mathit{Occ}(\mathit{Spreads}(A, H_1, H_2))(1)$ and its effect $\mathit{Occ}(\mathit{Fire}(A, H_2))(2)$ unsupported. It follows by minimization of occurrences at time 2 that $\neg\mathsf{T}\mathit{Occ}(\mathit{Fire}(A, H_2))(2)$ is true at w'. So it follows at w, as above, that the occurrence of the $\mathit{Spreads}(A, H_1, H_2)$ event at time 1 is the *Cause* of the $\mathit{Fire}(A, H_2)$ event at time 2.

Similarly, it follows at w that the occurrence of the $\mathit{Spreads}(A, H_4, H_3)$ event at time 1 is the *Cause* of the $\mathit{Fire}(A, H_3)$ event at time 2.

It follows from the maximization of inertia atoms at time 1 and Axiom (6) that $\neg\mathit{Burnt}(H_1)(2)$ is true at w. As the precondition of $\mathit{Occ}(\mathit{Fire}(A, H_2)(2)$ is true, it follows by maximization of successes at time 2 that the event succeeds, and consequently that its effect $\mathit{Burnt}(H_2)(3)$ is true at w; axioms (1), (12) and (13). Reasoning similar to that above shows that the $\mathit{Fire}(A, H_2)$ event is the *Cause* of H_2 being burnt.

The first conjunct to be proved now follows by Axiom (11).

The second conjunct is established by reasoning similar to that above. Let w' be the $\prec_C^{\Theta}$-closest $\neg\mathsf{T}\mathit{Occ}(\mathit{Fire}(A, H_1))(1)$-world below w; the support set for w', B', being obtained by removing $\mathit{Occ}(\mathit{Fire}(A, H_1)(1)$ from B. At w' we use event-theory reasoning, as above, to establish that fire B spreads to H_3 at time 2 and to H_2 at time 3, leaving H_2 burnt at time 4. Establishing a chain of unique *SCause*s is straightforward. Establishing that each event in this chain is contextually necessary is done by reasoning contrafactually in the $\prec_C^{\Theta}$-closest $\neg\mathsf{T}\mathit{Occ}(\mathit{Fire}(B, H_4))(1)$-world w'' below w'; whose support set is obtained by removing $\mathit{Occ}(\mathit{Fire}(B, H_4))(1)$ from B'. □

Example 2. van Fraassen's military-trumping example (Lewis 2000) can be represented by adding axioms (19)-(22) from Table 5. Soldier x can order soldier y to do e at time t iff x outranks y and anyone else who gives y an order at t; Axiom (19). Orders are normally carried out without question; Axiom (22). Soldier S_1 outranks soldier S_2, who in turn outranks soldier S_3; Axiom (21). Finally, S_1 and S_2 simultaneously order S_3 to advance; Axiom (22).

Let Θ be the causal theory with background laws $\Theta_L = \{(19), (20)\}$ and boundary conditions $\Theta_B = \{(21), (22)\}$. Then

$\Theta_2 \mid\!\approx_C \mathit{Cause}(\mathit{Occ}(\mathit{Ord}(S_1, S_3, \mathit{Adv}(S_3)))(1), \mathit{Occ}(\mathit{Adv}(S_3))(2)).$

Proof. Let M be the causal model for Θ with event pref-

erence relation $\prec_E^M$ and accessibility relation $\prec_C^\Theta$. There is a single E-preferred Θ-world, w, in M; Θ is thus deterministic. By maximization of successes (at w) at time 1, and axioms (1), (9), (19), (20) and (22), we conclude that the *Ord*$(S_1, S_3, \textit{Adv}(S_3))$ event succeeds at time 1, and that consequently it is an *SCause* of the occurrence of *Adv*(S_3) time 2. It also follows, from axioms (1), (19) and (21), that *Ord*$(S_2, S_3, \textit{Adv}(S_3))$ fails at time 1; because S_1 outranks, and so trumps, S_2. So it follows from the definition of *SCause* and the minimization of occurrences at time 1, that the *Ord*$(S_1, S_3, \textit{Adv}(S_3))$ event is the unique *SCause* of the subsequent *Adv*(S_3) event.

The support set for w, B, consists of the literals in axioms (21) and (22). It follows from Proposition 2 that removing *Occ*(*Ord*$(S_1, S_3, \textit{Adv}(S_3)))(1)$ from S, gives the support set, S', of the $\prec_C^\Theta$-closest $\neg\mathsf{T}$*Occ*(*Ord*$(S_1, S_3,$ *Adv*$(S_3)))(1)$-world, w' below w. Removing S_1's order makes the preconditions of *Occ*(*Ord*$(S_2, S_3, \textit{Adv}(S_3))(1)$ true at w'; Axiom (19). So it follows by maximization of successes (at w') that S_2's order succeeds at w', and consequently that this event is an *SCause* of the occurrence of *Adv*(S_3) at time 2. Its uniqueness as an *SCause* is established as above.

Removing *Occ*(*Ord*$(S_2, S_3, \textit{Adv}(S_3)))(1)$ from S' gives support set S'' and closest $\neg\mathsf{T}$*Occ*(*Ord*$(S_2, S_3,$ *Adv*$(S_3)))(1)$-world w'' below w'; Proposition 2. It follows by minimization of events (at w'') at time 1 that $\neg\mathsf{T}$*Occ*(*Adv*$(S_3))(2)$ is true at w''.

So $\neg\mathsf{T}$*Ord*$(S_2, S_3, \textit{Adv}(S_3))(1) \Downarrow \neg\mathsf{T}$*Occ*(*Adv*$(S_3))(2)$ is true at w'. It follows by Axiom (10) that *Occ*(*Ord*$(S_2,$ $S_3), \textit{Adv}(S_3))(1)$ is the *Cause* of *Occ*(*Adv*$(S_3))(2)$ at w'.

Therefore $\neg\mathsf{T}$*Ord*$(S_1, S_3, \textit{Adv}(S_3)) \Downarrow$ *Cause*(*Ord*$(S_2,$ $S_3, \textit{Adv}(S_3))(1), \textit{Occ}(\textit{Adv}(S_3))(2))$ is true at w. And so it follows by Axiom (10) that *Occ*(*Ord*$(S_1, S_3, \textit{Adv}(S_3)))(1)$ is the *Cause* of *Occ*(*Adv*$(S_3))(2)$ at w. □

Concluding Remarks

The proposed definition of causation is concerned with actual causation. This is naturally combined with the conditionals of $\mathcal{CL}$ to represent counterfactual causation; as illustrated by Example 1. As suggested in (Bell 2001a), complefactuals can also be used to generate explanations. Explanation can now be defined as follows

$$\forall \epsilon, \phi, \psi(\textit{Expl}(\epsilon, \phi, \psi) \equiv ((\epsilon \wedge \phi) \Uparrow \textit{Causes}(\epsilon, \psi))).$$

Thus occurrence ϵ together with conditions ϕ explain ψ at actual world w iff ϵ causes ψ at all closest $\epsilon \wedge \phi$-worlds above w. These ideas will be developed further in future work.

Future work will also include a more extensive empirical evaluation of the proposed theory of causation and a comparison with related work, particularly that of Halpern and Pearl (2001).

Finally, Proposition 2 suggests that it may be possible to extend the model-building implementation of primary events (White, Bell, & Hodges 1998) in order to implement the evaluation of causal counterfactuals. For example, in the simple case in which occurrence ϵ is not true at actual world w constructed by the algorithm for causal theory Θ, the closest ϵ-worlds above w can be generated simply by running the algorithm on $\Theta \cup \{\epsilon\}$.

References

Bell, J. 2000. Primary and secondary events. Linköping Electronic Articles: www.ida.liu.se/ext/etai/rac/. Subsequent version (2001): www.dcs.qmul.ac.uk/˜jb.

Bell, J. 2001a. Causal counterfactuals. Working Notes of Common Sense 2001: www.cs.nyu.edu/cs/faculty/davise/commonsense01/.

Bell, J. 2001b. Simultaneous events: Conflicts and preferences. In Benferhat, S., and Besnard, P., eds., *Proc. EC-SQARU 2001*, LNAI 2143, 714–725. Berlin: Springer.

Bell, J. 2003. A common sense theory of causation. In Blackburn, P.; Ghidini, C.; Turner, R. M.; and Giunchiglia, F., eds., *Proc. Context 2003*, LNAI 2680, 40–53. Berlin: Springer.

Fikes, R. E., and Nilsson, N. J. 1971. STRIPS, a new approach to the application of theorem proving to problem solving. *Artificial Intelligence* 2:189–208.

Gärdenfors, P. 1988. *Knowledge in Flux*. Cambridge Mass.: MIT Press.

Halpern, J., and Pearl, J. 2001. Causes and explanations: A structural-model approach. Part I: Causes. In *Proc. UAI 2001, 17th Conf. on Uncertainty in AI*, 194–202.

Hume, D. 1975. *Enquiry Concerning Human Understanding*. Oxford University Press. Reprinted from the 1777 edition.

Kleene, S. C. 1952. *Introduction to Metamathematics*. Amsterdam: North-Holland.

Lewis, D. 1973. Causation. *Journal of Philosophy* 70:180–191.

Lewis, D. 1986. *Philosophical Papers, Volume II*. Oxford University Press.

Lewis, D. 2000. Causation as influence. *The Journal of Philosophy* 97(4):182–197.

Schaffer, J. 2000. Trumping preemption. *The Journal of Philosophy* 97(4):165–181.

Shoham, Y. 1988. *Reasoning About Change*. Cambridge Mass.: MIT Press.

White, G.; Bell, J.; and Hodges, W. 1998. Building models of prediction theories. In Cohn, A.; Schubert, L.; and Shapiro, S., eds., *Proc. KR'98*, 557–568. San Francisco: Morgan Kaufmann.

Wright, R. 1988. Causation, responsibility, risk, probability, naked statistics, and proof: Pruning the bramble bush by clarifying the concepts. *Iowa Law Review* 73:1001–1077.