ELSEVIER

Artificial Intelligence 111 (1999) 73–130

Artificial Intelligence

www.elsevier.com/locate/artint

A commonsense language for reasoning about causation and rational action

Charles L. Ortiz Jr. [1]

SRI International, AI Center, EJ278, 333 Ravenswood Ave., Menlo Park, CA 94025, USA

Received 14 November 1997; received in revised form 14 January 1999

Abstract

Commonsense causal discourse requires a language with which to express varying degrees of causal connectedness. This paper presents a commonsense language for reasoning about action and causation whose semantics is expressed by way of counterfactuals. Causal relations are analyzed along several dimensions including notions of resource consumption, degree of responsibility, instrumentality, and degree of causal contribution. Grounding the semantics in a level of counterfactual reasoning is shown to play an important role in constraining the set of allowable event descriptions instantiating reports expressed by any of the relations in the language. These ideas are also applied to a causal analysis of rational action: by adopting an explanatory stance, one can characterize action through descriptions that refer to causal connections between mental states and actions. Such a causal analysis resolves some well-known difficulties in correctly ascribing agency and intentionality. Finally, an implementation is described—used to motivate and refine the theory—in which queries involving causal relations between the activities of agents engaged in purposeful behavior within a microworld can be posed.

Keywords: Causation; Reasoning about action; Agency; Commonsense reasoning

1. Identifying an event's causal role

The occurrence of an event is not an isolated matter. An event owes its existence to other events which causally precede it; an event's presence is, in turn, felt by certain collections of subsequent events. We attach descriptions to events within such chains and, if we wish to adopt an explanatory posture towards them, attempt to draw the connections, causal and otherwise, that identify *roles* among such events; the elaboration of such connections serving to further describe each individual event.

[1] Email: ortiz@ai.sri.com.

0004-3702/99/$ – see front matter
PII: S0004-3702(99)00041-7

The multiplicity of potential descriptions for a singular event combined with the generality with which we wish to endow our causal knowledge—which, it would seem, should range over *types* of events for the sake of achieving such generality—both conspire to complicate this causal attribution task. Let us consider how this might be so. Beginning with a world description, $\mathcal{WD}$, of facts about the world, our *a priori* knowledge also includes a body of law-like knowledge, $\mathcal{L}$, relating, for example, preconditions and effects to individual event-types, as well as other non-causal knowledge, Δ. Whereas $\mathcal{L}$ might include knowledge such as, "if block X is clear then after a $puton(Y, X)$ action, block Y will be situated on block X", the knowledge contained in Δ will instead contain knowledge such as "a block is clear just in case there is no block on top of it". Expressions contained in $\mathcal{L}$ will be referred to as *nomic expressions*.

To get a feeling for what is meant by causal role consider the following example in which one is filling a tank with water through some pipe, *In*, and one is also removing some of the water but at a much slower rate through a pipe, *Out*. This scenario supports the following causal report: "removing the water through pipe *Out* (event α) is slowing the rate at which the tank fills (event β)". Here, if one was trying to fill the tank, one might say that α contributed in a negative fashion to β's outcome. Assigning a role for an event is often aided by reference to some commonsense causal language whose elements include terms such as *causes, prevents, maintains, lets, helps*, and *hinders*; each of these reflect our intuitions regarding various sorts of dependencies between actions. In the above example, we might say that α is preventing the tank from filling up more quickly.

The task of causal attribution is complicated by two factors. First, there is the problem of how to avoid pointing to what will be called a *spurious connection*: a pair of events that are temporally adjacent but which are causally independent. In the above example, if one is talking on the phone at the same time the tank is filling then the action of talking on the phone may have absolutely nothing to do with β. In particular, it is well known that one cannot determine a causal connection between the occurrence of some α and some β by simply checking the truth of some material implication between formulas referring to the occurrence of each of the events [62,64]. A second difficulty stems from the inherent opacity of causal reports and the potentially unlimited number of descriptions with which one can refer to an event; whereas some pairs of alternate descriptions may represent genuine causal connections others might not: for example, while *John's taking highway 101 to work caused him to be late for the meeting* might represent a true causal report, *John's driving to work in his buick caused him to be late for the meeting* might not. One cannot, therefore, simply read off the "connections" between event descriptions specified in $\mathcal{L}$: there are simply too many of them; many of these descriptions must be formed by way of knowledge in Δ. This will be referred to here as the *event subsumption problem*.

1.1. Approach

This paper examines the application of counterfactual reasoning to the problem of identifying the role that an event played in some broad nexus of events. The focus is on:

(1) developing a language of commonsense causal relations such as *causes, prevents, enables, forces, lets, helps, hinders*, and *maintain* that are useful to this task,

(2) examining how the appropriate event description to instantiate the chosen causal relation can be chosen, and
(3) applying these ideas to the causal analysis of rational action.

Categorizing rational action is often simplified by adopting an *explanatory stance* involving the identification of causal connections between mental states and actions as well as method-of relations between complex actions and simpler, more primitive actions.

The choice of counterfactuals as a tool is motivated by the intuition that in considering what role some α might have played in some β one should "imagine" the consequences to β of α not occurring. For example, one might justify the statement that

(1) Mary helped Peter cut the grass.

by counterfactually arguing that Mary performed some action—say, cutting the grass in the backyard—which, if she had not performed, would have resulted in Peter taking longer to finish. In evaluating such a counterfactual one must naturally imagine that everything else remained the same, that is, that Peter did not instead work faster. Counterfactuals are also of use in the task of choosing the appropriate event description in such attributions. If Peter happened to cut the grass in a sloppy manner then it may be inappropriate to point to Mary as having any responsibility or role in such an eventuality: assuming Mary did a good job on the backyard, even if Mary had not cut the grass in the backyard, it would not have had any effect on the outcome of Peter's work; hence, no support would be provided for the report:

(2) Mary helped Peter cut the grass in a sloppy manner.

which pairs Mary's actions with the event described as *cut the grass in a sloppy manner*.

Counterfactuals are useful in expressing many important causal relations: relations such as preventions as well as acts of, for example, maintaining, letting, and coercion all call for a counterfactual analysis. For example, one natural way to define a prevention of some α is to identify it with some actual event, β, whose absence of occurrence might have resulted in α. Similarly, one natural way to verify that a causal connection did indeed exist between two actions is to say that α caused β just in case both α and β occurred and if α had not occurred β would not have occurred.

1.2. Comparison to related work

Research on causation in AI can be roughly divided into three areas: work in planning or reasoning about action, explanation, and linguistic analysis. The present work falls primarily in the area of explanation with a focus on the event subsumption problem, a problem which has not been previously explored. In contrast, work in planning and linguistics has focussed primarily on the representation of nomic expressions and their use. The representation of the numerous causal primitives discussed in this paper, however, are not without use in planning tasks. Agents collaborating on a task might find it useful to explain the behaviors of other agents: for example, whether another agent accidently or intentionally α-ed.

Some of the early work relevant to the topic of this paper is that of Rieger [60] who investigated a broad set of causal primitives involving notions such as "one-shot" causation

and enablement. For example, one-shot causes are distinguished by a causing event which initiates a process which then continues independently of the causing event:

(3) Hitting the cue ball caused it to strike the nine ball and put the latter in motion.

After the event of hitting the cue ball, everything that occurs is independent of the initial causing event. Subsequently, Schank [61] and his students developed a theory called *conceptual dependency* which included primitives such as *p-trans* (for physical transfer), *a-trans* (for transfer of possession), *propel*, *grasp*, and a host of others. It was not until McDermott's work on branching time [50] that these ideas were placed on a firm temporal foundation: causal relations were explored in terms of the possibilities or future "chronicles" that they opened or closed. For example, McDermott defined a predicate called ecause which is true of two events, α and β if α is followed by β in every possible future unless some fact, p, becomes false at some point. The fact p can be seen as a sort of enabling fact of the type suggested by Rieger [60]. McDermott's approach, however, was not based on counterfactuals and would therefore be susceptible to the following problem: α might depend on some previous event or fact and therefore the assumption of α not occurring might be inconsistent with those facts or events in the past. Allen [1] also makes use of a causal connective but does not supply a semantics for it.

Shoham [62] focussed on other related problems: the frame problem and a theory of causal relations in which the problem of spurious causal connection was avoided by restricting causal influence to inferences involving only causal rules. Shoham also presented a semantics for the primitives *causes*, *enables*, and *prevents*. Neither McDermott's nor Shoham's work was concerned with the event subsumption problem. There has been much work since then on the frame problem, but that question is orthogonal to the aims of this paper. Finally, a large body of important work has involved the study of *qualitative reasoning*: that is, the study of systems that reason about and explain the behavior of physical systems in qualitative rather than quantitative terms [74]. It is a pity that such work has often been pursued independently of formal accounts of theories of change, though Eiter and Gottlob [25] briefly discuss belief revision in the qualitative reasoning domain.

These questions have not only been addressed by the field of AI, but also in linguistics and philosophy. Talmy [66,67] put forward a theory of what he termed "force dynamics"—a semantic category that deals with the interaction of objects by way of forces exerted by those objects; the resulting theory represents a rather fine-grained account of a number of types of causation, such as causing, letting, hindering, maintaining, and forcing. The notion of a "force" is much broader than that found in physics; it is meant to also include, for example, psychological elements through which one can analyze instances of actions such as refraining and resisting. The theory assigns the roles of agonist (the focal force) and antagonist (the opposer) to entities involved in the force interaction, and then expresses force relationships between those entities by way of an informal diagrammatic representation. A problem with such an approach, however, is that it is often difficult to see how one can assign forces in a uniform way. From an AI perspective it seems preferable to use notions of persistence and temporal projection from which those force relationships could be derived. In addition, the importation of force dynamics to psychological categories is not based on any well-developed theory of rational behavior; that is, the relationships between beliefs, desires, intentions (BDIs) and action are not explicated.

A number of linguists have undertaken a semantic study of causal primitives. In Chapter 7 of Jackendoff [38], Talmy's approach is applied to linguistic analysis. For example,

Harry forced Sam to go away

(Example 17, p. 131) is analyzed as (using instead a logical-style notation):

$$CAUSE(Harry, (go(Sam, Away), Aff(Sam))) \wedge Aff(Harry, Sam)$$

where *Aff* is a primitive meant to assign the role of agonist to Sam; that is, the person on whom the force is applied. $Aff(Harry, Sam)$ specifies that Harry is applying the force and is trying to bring about $go(Sam, Away)$. However, this representation does not distinguish the above from instances of *causing Sam to go away* nor does it specify exactly under what circumstances Sam can be said to be resisting Harry's influence, a seemingly important characteristic of instances of being forced.

The work on causation in philosophy is too numerous to review in its entirety here. Noteworthy with respect to the focus of this paper is Von Wright's [73] model of branching time which was used to define an agent-centered version of causation. Brand [12] also explored a number of causal relations and act types in his work. Finally, Lewis [45] and Dowty [24] both explored a counterfactual semantics for causation. The interested reader is referred to the excellent volume by Sosa [64] that presents many varied philosophical viewpoints on causation.

1.3. Outline of paper

The remainder of this paper is organized in the following way. First, a representation language in which the subsequent analysis is carried out is briefly discussed (the syntax and semantics of the language is described in detail in Appendix A). This leads to a discussion of the semantics of causation. A number of causal and non-causal relations are then examined with the help of this machinery. A micro-world implementation that served to test and refine the theory is described. Finally, the last section summarizes the various relations examined.

2. Representation language

The representation language, Hypothetical Logic ($\mathcal{HL}$), was developed to investigate the application of counterfactual reasoning to causation [58]. $\mathcal{HL}$ is a sorted modal first-order language with sorts for events, times, fluents (properties of the world that change with time), and objects. The language contains two predicates: $occurs(e, t)$ reports the occurrence of event type e at time t, and $holds(f, t)$ reports that fluent f is true at time t; time constants range over the integers and the truth of a formula, ϕ, is given relative to some world, w, and is written $w \models holds(\phi, t)$. The duration of an event is specified in the description of the event by way of an event-type constructor which creates complex event-types from simpler ones: for example, $pickup@manner(slowly)@dur(10)$ might name the event type of picking up an object slowly, that event having taken 10 time

units. Additional axioms allow one to draw adverb-dropping inferences: from the fact that *pickup@manner*(*slowly*)@*dur*(10) occurred to the fact that *pickup* occurred. The language also contains a possibility operator: $holds(\Diamond\phi, t)$ means that there is some physically possible world in which ϕ holds. An agent i's belief at time t in some ϕ is expressed as $holds(Bel(i, \phi), t)$, where ϕ can be a temporal term written in one of the functional forms $Holds(\psi, t')$ or $Occurs(\alpha, t')$.[2] A statement of the form $\Box_W\phi$ will mean that ϕ is true in all possible worlds. Finally, the connective, $>$, standing for counterfactual dependence will be employed in some of the definitions; the semantics of counterfactual dependence is discussed at length elsewhere [57,58]. Statements of the form $\phi > \psi$ are meant to express the fact that if ϕ had been (instead) true, then so would have ψ.[3] In order to deal with statements involving individuals that might not exist in some worlds, an existence function is added: $w \models holds(E(X), t)$ means that X exists in world w at time t (see Appendix A).

Complex event types can be constructed not only by way of the event-type constructor but also through a number of operators that one normally sees in dynamic logic:

$occurs(\alpha; \beta, t)$: α is followed by β;

$occurs(\alpha^*, t)$: α occurs zero or more times;

$occurs(\alpha \cap \beta, t)$: both α and β occur at t;

$occurs(\alpha \cup \beta, t)$: either α or β occurs at t;

$occurs(\phi?, t)$: is true if $holds(\phi, t)$ is true.

From these operators one can then construct expressions such as:

$occurs(\text{WHILE } \phi \text{ DO } \alpha, t)$

which reports the repeated occurrence of α just as long as the condition ϕ holds, and *if-then-else*-statements:[4]

$occurs(\text{IF } \phi \text{ THEN } \alpha \text{ ELSE } \beta, t)$

which reports the occurrence of α if the condition ϕ holds and the occurrence of β otherwise.

The details of the syntax and semantics of $\mathcal{HL}$ are described in Appendix A.

3. A counterfactual analysis of causation

Counterfactual dependency represents a necessary condition to support the conclusion of the existence of a causal connection, but does not represent a sufficient one [58]. In this section, two sufficiency conditions are placed on the definition of causation in

[2] Notice the use of upper case.

[3] In the remainder of this paper, upper case symbols will stand for constants and lower case for variables. In addition, formulas will be assumed to be universally quantified.

[4] In the remainder of this paper, upper case symbols will stand for constants and lower case for variables. In addition, $\alpha, \beta, \gamma, \delta$ will stand for event types whereas ϕ, ψ, ξ will stand for fluents, and also formulas will be assumed to be universally quantified.

addition to that necessary condition rooted in counterfactual dependency, that is to that condition which states that $occurs(\alpha, t_1)$ causes $occurs(\beta, t_2)$ just in case both α and β actually occurred but if α had not occurred then neither would β have occurred. These two conditions involve restricting the causal relation to exclude cases in which the antecedent event is either *part of* the consequent event or represented a *method* for the consequent event. This observation owes a debt to certain criticisms leveled by Kim [40] on Lewis' program of counterfactuals. Kim points out that some pairs of events that depend counterfactually on each other are not causally related. Take, for example,

(4a) I played the C chord. If I had not played the E note
I would not have played the C chord.

(4b) I opened the door by pulling it. If I had not pulled the door
I would not have opened it.

Both counterfactuals are true. However, it would be incorrect to assent to the following causal reports.

(5a) #My playing the E note caused me to play the C chord.

(5b) #My pulling the door caused me to open it.

We would rather say that, in the first case, my playing the E note was a *part* of my playing the C chord; in the second case we would say that my pulling the door was the *method* I employed to open the door. These cases are handled as follows. First define the transitive closure of counterfactual dependence, $>^*$. This is necessary as counterfactuals are inherently nonmonotonic and, hence, not generally transitive [28].

Definition 3.1 (*Transitive closure of counterfactual dependence*). The transitive closure, $\overset{*}{>}$, of $>$ is defined by:

$\models \phi \overset{*}{>} \psi$ iff

either $w \models \phi > \psi$

or there is some ξ such that, $w \models \phi > \xi$ and $w \models \xi \overset{*}{>} \psi$.

The following will be used to block problematic cases of *pre-emption* that have been used as arguments in the literature against a counterfactual analysis of causation. Horwich observes [36, pp. 211–212]:

> For example, Smith's shooting a man pre-empts Blogg's shooting him if Bloggs is frightened off before firing by the sound of Smith's gun. If C's causing E pre-empts G's causing E, then, on the face of it, E is not counterfactually dependent on C because, even if C hadn't occurred, E would have been caused by G instead. Thus pre-emption might seem to present a problem for the counterfactual theory of causation.

The following definition for the symbol $\triangleright$ restricts $\overset{*}{>}$ to hold between a condition that represents the *strongest antecedent* condition counterfactually related to another. This definition is defined meta-theoretically, indicated by the $\equiv_{def}$.

Definition 3.2 (*Strongest antecedent condition*).

$$
\begin{aligned}
&occurs(\alpha, t_1) \rhd occurs(\beta, t_2) \equiv_{def} \\
&\quad \exists\gamma_1 \cdots \exists\gamma_n.[\Box_W[occurs(\alpha, t_1) \supset occurs(\gamma_1, t_1) \vee \cdots \vee occurs(\gamma_n, t_1)] \\
&\qquad \wedge (occurs(\gamma_1, t_1) \vee \cdots \vee occurs(\gamma_n, t_1) \overset{*}{>} occurs(\beta, t_2)) \\
&\qquad \wedge (\neg(occurs(\gamma_1, t_1) \vee \cdots \vee occurs(\gamma_n, t_1)) \overset{*}{>} \neg occurs(\beta, t_2))] \\
&\qquad \wedge \neg\exists\gamma_1' \cdots \exists\gamma_m'.[\Box_W((occurs(\alpha, t_1) \supset occurs(\gamma_1', t_1) \vee \cdots \vee occurs(\gamma_m', t_1)) \\
&\qquad \wedge (occurs(\gamma_1', t_1) \vee \cdots \vee occurs(\gamma_m', t_1) \supset occurs(\gamma_1, t_1) \vee \cdots \vee occurs(\gamma_n, t_1)) \\
&\qquad \wedge ((occurs(\gamma_1', t_1) \vee \cdots \vee occurs(\gamma_m', t_1)) \overset{*}{>} occurs(\beta, t_2) \\
&\qquad \wedge (\neg(occurs(\gamma_1', t_1) \vee \cdots \vee occurs(\gamma_m', t_1)) \overset{*}{>} \neg occurs(\beta, t_2))))]
\end{aligned}
$$

That is,

$$occurs(\gamma_1, t_1) \vee \cdots \vee occurs(\gamma_n, t_1)$$

is the strongest condition/action whose occurrence is counterfactually related to β such that the occurrence α entails the occurrence of the γ. As usual, α, β and the γ's can be test actions.

For example, suppose that there are two ways to win a particular chess game (β): by taking a certain rook or by advancing a certain pawn. Then, while the negation of each individual fact is not counterfactually related to not winning the game, the disjunction "taking the rook or advancing the pawn" (γ) is. Furthermore, if one of these actually occurred—say, taking the rook(α)—then we can report (see below) that taking the rook caused the win. This definition makes reference to a sort of *disjunctive event*. The existence of such an event type is problematic [47]. However, this axiom does not make any claims about the ontological status of such events. It is used rather as a means of getting around the pre-emption problem: since α must be part of the strongest antecedent condition, the definition will rule out cases in which α might not be counterfactually related to β simply because if α had not obtained, it might have pre-empted some other cause that would have resulted in β. Such cases are ruled out because we have, in effect, restricted the counterfactual dependency in the definition for causation to *classes* of events.

Causation can then be defined as follows.

Definition 3.3 (*Semantics for causation*).

$$
\begin{aligned}
&occurs(\alpha, t_1) \textbf{ causes } occurs(\beta, t_2) \equiv_{def} \\
&\quad \alpha \neq \beta \vee t_1 \neq t_2 \\
&\quad \wedge \neg holds(augmentations(\alpha, \beta), t) \\
&\quad \wedge occurs(\alpha, t_1) \textbf{ instrumental } occurs(\beta, t_2) \\
&\quad \wedge \neg(occurs(\alpha, t_1) \textbf{ method } occurs(\beta, t_2)) \\
&\quad \wedge \neg holds(part\_of(\alpha, \beta), t) \\
&\quad \wedge occurs(\alpha, t_1) \rhd occurs(\beta, t_2)
\end{aligned}
$$

In words, the fact that α occurred caused it to be the case that β occurred just in case the following conditions are met:

(1) either α and β represent distinct act types or they occurred at different times;
(2) neither represent augmented descriptions of the other (this rules out cases such as $\alpha = e@p@q$ and $\beta = e@p$—see below);
(3) α is instrumental in the performance of β (this rules out cases in which α did not cause β simply because it did not prevent β which would have occurred on its own, and so the counterfactual dependence in the last clause will hold—see below);
(4) α must not represent a method for β (see below);
(5) α is not a part of β: this precludes incorrectly classifying cases of one event being a part of a sequence as genuinely causal; and finally
(6) the occurrence and nonoccurrence of α and β, respectively, are each counterfactually related in terms of the strongest antecedent condition, for the reasons already discussed.

Examples such as the following are interesting from the perspective of case (4):

(6) Picking up the heavy sofa caused him to hurt his arm.

where "Picking up the heavy sofa" also represented a method for "hurting his arm". It is not clear how such cases should be handled: $\mathcal{HL}$ would require this last sort of example to be reported as "Picking up the heavy sofa caused his arm to become hurt". Whether the example in (6) can be handled by appealing to some sort of intentionality is not clear. Nevertheless, the above formulation seems to correctly handle a great number of useful cases.

The notion of instrumentality employed here is somewhat similar to that suggested by Bennett [10]. He suggests identifying cases of an agent's behavior as instrumental to some consequence, ϕ, just in case the number of basic movements the agent could have performed that resulted in ϕ is strictly smaller than the number of movements that would have resulted in $\neg\phi$. The following definition instead appeals to the existence predicate/function.

Definition 3.4 (*Instrumentality*).

$$\begin{aligned}&\models \mathit{occurs}(\alpha, t_1)\ \textbf{instrumental}\ \mathit{occurs}(\beta, t_2) \equiv_{def}\\ &\quad(\exists a.\mathit{holds}(\mathit{agt}(\alpha, a), t_1)\\ &\quad\wedge \mathit{occurs}(\beta, t_2) \wedge [\mathit{holds}(\neg E(a), t_1) \overset{*}{>} \neg\mathit{occurs}(\beta, t_2)])\end{aligned}$$

That is, β must occur and if the agent of α had not existed then β would not have occurred.[5] The requirement that β occur is necessary to handle cases of "letting" discussed later.

It is also useful to extend the definition for causation to report cases in which some event "actually" caused another. We have:

[5] See Rationality Postulate A.1 in the appendix.

Definition 3.5 (*Actually caused*).

$$\models \alpha \textbf{ caused } \beta \equiv_{def} (\alpha \textbf{ causes } \beta) \wedge \alpha \wedge \beta$$

Shoham [62] examined a number of properties of causation: transitivity, asymmetry, and ireflexivity. In my definition, asymmetry falls out of the preferences already discussed and, unlike Shoham's theory does not require the stipulation of temporal precedence between the events involved. As such, the semantics presented here can handle cases of simultaneous causation such as, *the block scratched the table as it was dragged across the surface*. Transitivity is built into the definition of $>^*$ and ireflexivity falls out of the requirement that the events be distinct.

The property of transitivity is somewhat controversial, however, and deserves comment. For example [34]:

> "... the cause of the motor accident may be the icy condition of the road but it would be odd to cite the cold as the cause of the accident."

Similarly, from the first pair of statements one should not conclude the last pair [51]:

(7a) When John left, Sue cried. When Sue cried, her mother got upset.

(7b) When John left, Sue's mother got upset.

These examples seem to suggest that causality is not transitive. However, Thomson [68, pp. 61–62] and Lewis [46, pp. 214–215] blame this on a lack of distinction between "a cause" and "the cause". In the example of Hort and Honoré a number of factors contributed to causing the accident: the icy condition, the cold, the attention of the driver, and so forth. Each of those conditions constituted "a cause", however, "the cause" depends on one's explanatory goals: in this case, the icy condition represents the most salient cause. This points to one of the differences between causal reasoning in support of planning and causal reasoning in support of explanation.[6] The former is not concerned with the distinction between "a cause" and "the cause"; but rather in predicting the future in terms of some set of goals. These subtle distinctions, however, are important for explanatory tasks and appear prominently in, for example, the legal arguments examined by Hart and Honoré. In this paper, it will be assumed that causation is transitive and that the explanation for the compelling examples of Hart and Honoré as well as Moens and Steedman are to be found at a pragmatic level which is responsible for identifying *the* salient cause among a number of causal factors [57].

4. Grounding actions: The method-of relation

Turning now to the notion of a *method* for an action, it is useful to first define the notion of *augmentation*: one action description augments another just in case one represents a more detailed description of the other. Formally,

[6] The other difference has to do with the event subsumption problem already discussed.

Definition 4.1 (*Augmentation of events*).

$$\begin{aligned}&\models \mathit{holds}(\mathit{augmentations}(\alpha,\beta),t)\equiv\\&\quad[\mathit{holds}(\mathit{augments}(\alpha,\beta)\\&\quad\vee \mathit{augments}(\beta,\alpha),t)]\end{aligned}$$

where,

$$\begin{aligned}&\models \mathit{holds}(\mathit{augments}(\alpha@\mathit{mods},\alpha@\mathit{mods}'),t)\equiv\\&\quad\mathit{holds}(\mathit{sub\_type}(\mathit{mods}',\mathit{mods}),t)\end{aligned}$$

What will be called here the method-of relation between actions is very similar to the generation relation of Goldman [29]. The following, somewhat tired example, should illustrate the intuition behind generation. Consider the act of flipping some particular light switch at some particular time; that action is said to *generate* the act of turning on that light at that time just as long as some set of conditions are in force: in this case, the switch and the light must be connected and functioning in the appropriate manner. In such a case one would say that one turned on the light *by* flipping the switch.

Definition 4.2 (*Method of performing an action*).

$$\begin{aligned}&\mathit{occurs}(\alpha,t)\ \textbf{method}\ \mathit{occurs}(\beta,t)\equiv_{\mathit{def}}\\&\quad\mathit{holds}(\alpha\neq\beta\\&\quad\wedge\exists a.\mathit{agt}(\alpha,a)\wedge\mathit{agt}(\beta,a)\\&\quad\wedge\mathit{duration}(\alpha)=\mathit{duration}(\beta),t)\\&\quad\wedge\neg\mathit{augmentations}(\alpha,\beta)\\&\quad\wedge\mathit{occurs}(\alpha,t)\rhd\mathit{occurs}(\beta,t)\end{aligned}$$

That is, some α represents a method for some β just in case they are distinct, performed by the same agent, and over the same time spans. In addition, β should not represent a more detailed description of α and α and β should be counterfactually related. Appropriate definitions for the *agent*, *time*, and *duration* functions are assumed.

This analysis—and also Goldmans'—takes the time of the ends or generated action to be contemporaneous with the means action. Hence, the turning on of the light occurs at exactly the same time as the flipping the switch action, even though a very brief span of time separates the two. The problem of the "time of action" is recognized as a notoriously difficult one, however [11,26,68]. For example, if someone is shot but dies an hour later, at what time did the killing take place? These philosophical questions are beyond the scope of this paper.

The use of $\alpha\rhd\beta$ instead of $\overset{*}{>}$ in the definition blocks cases of pre-emption, just as in the causal case. Consider,

(8) Bobby won the game by taking his rook.

Given Bobby's commitment to winning, he would have chosen some other winning method if it existed. [7]

In the definition for method-of, the act-types are constrained to not represent augmentations of each other; this is done in order to treat separately cases which Goldman [29] referred to as *augmentation generation*. These are cases relating the italicized descriptions below in the manner indicated:

(9) ??John *ran quickly* by *running*.

This case of the *by*-locution does not sit comfortably with his other examples. These are handled here separately as adverb-dropping inferences along the lines of a Davidsonian approach [22].

Since the definition does not require that one identify some basic action that is related to the higher-level action (as, for example, Israel et al. [37] does), the definition correctly handles cases involving non-movement and negative actions:

(10a) He angered the teacher by not answering the question.

(10b) He kept the coffee from spilling by maintaining the cup steady.

(10c) He prevented the flood by closing the gates.

In the first example, if he *had* answered the question, he would not have angered the teacher; the event subsumption is therefore handled without having one's knowledge base include a statement that specifically relates the two act-types in the manner indicated. The reports (10b) and (10c) both illustrate that not every action can be equated with a cases of bringing about some condition; in fact, the world remains visibly the same even though "something is happening". In the second case, there need not be any observable change in the position of the cup. In the third, the action that is prevented never occurs. Finally, notice that Definition 4.2 leads to a relation that is transitive, anti-symmetric, and irreflexive. This is consistent with Goldman's observations and once again is discussed in more detail in the next section.

Given the above, the composite *by* action can now be defined as in the case of "actually caused".

Definition 4.3 (*The by-locution*).

$$\models \mathit{occurs}(\mathit{by}(\beta,\alpha),t) \equiv$$
$$\mathit{occurs}(\alpha,t) \wedge \mathit{occurs}(\beta,t) \wedge \mathit{occurs}(\alpha,t)\ \textbf{method}\ \mathit{occurs}(\beta,t)$$

That is, an agent performs some β by α-ing just in case both α and β occur over identical spans of time and, moreover, α represents a method for the performance of β.

The above definitions correctly handle the examples that distinguish causal from non-causal connections discussed earlier. There are other, more problematic cases, however.

[7] Thanks to Mark Steedman for this example. As we shall see in the next section, similar criticisms had been leveled on counterfactual treatments of generation by, for example, [14,30,59].

These were also pointed out by Kim [41]. He notes that whereas the first of the following pair is true the second is not.

(11a) If Socrates had not died at t, Xantippe would not have become a widow at t.

(11b) Socrates' death caused Xantippe to become a widow.

Kim believes that the second does not reflect a genuine causal connection. For one thing, Xantippe and Socrates were not located at the same place, hence a causal connection would appear to involve action at a distance; Kim is of the belief that a causal connection must involve some contiguous set of connections. Perhaps it would be more reasonable to state that

(12) Socrates' death resulted in Xantippe becoming a widow.

(See the discussion in [70].) In any case, under the present framework this sort of noncausal connection is difficult to handle. It might be best to deal with such examples by formalizing some notion of context so that our commonsense laws of change are related to more physical ones; in the end, then, one would require some sort of spatial adjacency between the objects involved in the causal relation; this requirement is lacking in example (11). Another possibility is to stipulate that causality in social systems is distinct from that in planning systems. Yet another approach is that of Lewis who sought to translate the above problem into one of event individuation: two events are related causally just in case they stand in a certain counterfactual dependence and those events are *distinct*. In the case of (11), he would argue that the events in question are the same. However, the problem of event identity, is a very difficult one although Lewis [47] provided some guidelines intended to assist one in resolving these questions in terms of a notion of *event essence* (see also Yagisawa [69] for a related discussion).

4.1. Negative actions

Many philosophers have debated the relative merits of admitting such an ontological category [10,12,68,72], arguing on the one hand that negative actions do exist and are to be distinguished by some generative pro-attitude or intention [72]. Others deny their existence, noting that such entities are usually referred to by way of imperfect nominals—either infinitival or gerundial nominals—("John tried not to spill the coffee" or "John's not attending the meeting caused resentment by his co-workers") [10]. As observed by Vendler [70], imperfect nominals pick out states of affairs and not events which, according to Vendler, are normally referred to by way of perfect or derived nominals ("The attendance at the meeting") [71]. Others note the apparent difficulty involved in locating negative events in space and time and offer that as evidence for their non-existence [68]. In psychology, some interesting work has involved studies of the biases that arise when people judge the losses brought about from an action versus the losses brought about through some omission [9].

This paper is not concerned with the resolution of these important philosophical questions, but rather on the advantages derived from reifying negative events to problems of explanation generation or planning. With respect to the latter, an agent interacting with

other agents may on occasion have a pressing need to recognize that another agent avoided some action, α, by performing some other action. With respect to the latter, an agent might have a need to explain to some other agent by what means that agent, for example, avoided α-ing.

4.1.1. Basic negative actions

As in the case of positive actions, a negative action for which there does not exist some other action which represented a method for its performance will be labeled here a *basic negative action*. These turn out to be rather problematic. In the simplest case, we have examples of refrains, discussed in a later section, in which an agent simply did not α when it could have. Another example of a basic negative action is *neglecting* to α when one should have: this obligation might be based either on rationality considerations—α-ing might have been the best choice for carrying out the agent's ends action—or on some deontic context which specifies what one should or can do.

No attempt will be made to catalog all of the different ways in which one could avoid, omit, neglect, refrain from, α-ing, where α is not done by way of some other action; such a task is beyond the scope of this paper. One possible strategy that would be consistent with the treatment so far discussed for complex negative actions involves adopting a sort of *explanatory stance*. If one adopts the view that actions stand in causal relation with an agent's mental state, then one can say that an agent not-α-ed when in some partial mental state R because if R had not held then the agent would have α-ed instead. For example, *He refrained from smoking because he saw the no smoking sign and did not wish to break the law. Had he not held these latter beliefs then he* would *have smoked*. Actions then become simply tuples where each action is grounded, as before, in either a basic movement or some partial mental state description (or, as in the case of accidents, etc., as a combination of the two).

5. Comparison to generation

In this section the definition of method-of formalized above is compared with Goldman's [29] notion of generation. The axiomatization that follows is modeled on work by Pollack in plan recognition [59] and Balkanski [7]. Goldman observed that generation is an antisymmetric, irreflexive, and transitive relation between act-tokens:

(13a) John signaled by waving but he did not wave by signaling.

(13b) John did not signal by signaling.

(13c) John signaled by waving. By signaling to the auctioneer, John placed a bid on the item being auctioned. Therefore, John placed a bid by waving.

Cases of transitivity are subject to the same sorts of pragmatic constraints that affect causal reports, as discussed earlier: for example, as one gets further "away" from the generated action, the generating action becomes less and less significant in the production of the generated action. Take (13c) one step further:

(14) ??John bought the painting by waving.

Goldman originally identified four types of generation: causal, conventional, simple, and augmentation. The following are typical examples of the first three:

(15a) He turned on the light by flipping the switch (causal generation: the flipping of the switch *caused* the light to come on but was not a cause of turning on the light).

(15b) He signaled by waving (which appeals to the common convention that waving can be interpreted as signaling, in certain circumstances).

(15c) Bill jumped 6′3″. John then out-jumped Bill by jumping 7′0″ (the proposition expressed by the first sentence describes the circumstances under which the second is true).

The following is a restatement of Pollack's formalization of Goldman's definition in $\mathcal{HL}$. Note that Pollack questioned the last two counterfactuals and did not include them in her formalization.

$$
\begin{aligned}
&\mathit{occurs}(\alpha, t)\ \textbf{generates}\ \mathit{occurs}(\beta, t) \equiv_{def} \\
&\quad \exists c \forall t'.\Box_W[\mathit{holds}(c, t') \wedge \mathit{occurs}(\alpha@\mathit{agt}(i)@\mathit{dur}(d), t') \\
&\qquad\qquad \supset \mathit{occurs}(\beta@\mathit{agt}(i)@\mathit{dur}(d), t')] && (1) \\
&\quad \wedge \Diamond_W \exists t'.[\mathit{occurs}(\alpha@\mathit{agt}(i)@\mathit{dur}(d), t') \wedge \neg \mathit{occurs}(\beta@\mathit{agt}(i)@\mathit{dur}(d), t')] && (2) \\
&\quad \wedge \Diamond_W \exists t'.[\mathit{holds}(c, t') \wedge \neg \mathit{occurs}\beta@\mathit{agt}(i)@\mathit{dur}(d), t')] && (3) \\
&\quad \wedge [\neg \mathit{occurs}(\alpha@\mathit{agt}(i)@\mathit{dur}(d), t) > \neg \mathit{occurs}(\beta@\mathit{agt}(i)@\mathit{dur}(d), t)] && (4) \\
&\quad \wedge [\mathit{holds}(\neg c, t) \wedge \mathit{occurs}(\alpha@\mathit{agt}(i)@\mathit{dur}(d), t) > \\
&\qquad \neg \mathit{occurs}(\beta@\mathit{agt}(i)@\mathit{dur}(d), t)] && (5) \\
&\quad \wedge \mathit{holds}(c, t) && (6)
\end{aligned}
$$

The first clause states that there must be some condition, c, such that whenever c holds and α occurs, β does as well. This clause is prefaced by the operator $\Box_W$. This corrects Pollack's formulation which instead included only the $\forall$ quantification over time instants: the problem is that, in the actual world, α might simply have never occurred so that the question of whether the resulting choice of c is proper becomes moot.

The condition c is referred to by Pollack as a *generating-enabling condition.* Notice that c cannot stand for a complete description of the world, otherwise all of the conditions would be trivially satisfied and any two acts that occurred would be related by generation; this possibility is formally disallowed by clause (3) which requires that c not be sufficient for β. What c can or cannot stand for is an open question, hence a difficult choice. Castaneda [14] has shown that there is always some trivial c that satisfies the above clauses, while Bennett [11] remarks that one must distinguish "honestly free-standing conditions and ones that are constructed out of α and β".

Implicit in the representation of (1)–(6) is the following: α and β are required to occur over an identical time span, t. This is meant to capture Goldman's stipulations: that neither α nor β be a temporal part of one another, that neither α nor β be subsequent to the other, and that α and β not be, what he terms, *co-temporal*. The latter is informally identified

by Goldman with any case in which one might be inclined to say that either α or β was performed "while also" performing the other. [8]

Clauses (2) and (3) are meant to capture a condition of Goldman's which states that neither α nor c alone should entail β; this precludes the possibility of α and β being identical. Once again, these are prefaced with the two modalities for the same reason mentioned earlier. Clauses (4) and (5) are both part of Goldman's original definition. Clause (4) says that if the agent had not done α then he would not have done β, while clause (5) says that if c had not obtained, then even though i α-ed, i would not have β-ed. Finally, clause (6) says that the condition c holds.

Clauses (4) and (5) were suggested by Goldman in order to handle what he referred to as branching acts. A proper treatment of branching acts was not central to either Pollack's or Balkanski's analysis. Branching acts are cases in which one act generates two other acts. To take one of his examples, playing the piano might generate putting Smith to sleep while simultaneously generating the awakening of someone named Brown. Clause (4) allows one to conclude that the two generated acts are not related to each other by generation. However, Pollack observed that the counterfactual, "if the piano had not been played Smith would not have gone to sleep" might be false because Smith might have fallen asleep anyway. This problem has already been addressed by restricting the counterfactual dependency to the strongest antecedent condition and also by limiting oneself to predictive counterfactuals where preference is given to the stability of the past [57,58]. Israel et al. [37] consider branching acts (though not under that name) suggesting an alternative analysis for them: that the circumstances under which one of the acts, say putting Smith to sleep, is generated by the playing of the piano are different than those which generate the other action. But then one can only have both cases of generation go through if one is willing to accept the fact that putting Smith to sleep also generated waking Brown up, which seems counterintuitive. [9]

Returning now to clause (5), that clause was actually included by Goldman in order to ensure that generation take the proper direction. To take his example, if c is chosen so that any case of an arm extension outside of a car window is equated with a signal, then the application of five blocks the possibility that the signal generated the arm extension. Pollack observed a potential problem: this inference might not be sanctioned *because* of the equivalence: "if these two acts were biconditionally related, then G [the agent] would do both, or do neither". But that depends on which particular theory of counterfactuals one adopts. For example, if one adopts a syntactic approach (as this paper does) in which *only* the particular formula in question—in this case, our c—is withdrawn from the database then there is no contradiction in having other axioms survive the counterfactual supposition, namely axioms of the form: "if muscle twitch of type X in arm then arm is extended", so that the agent can still extend his arm in the counterfactual world.

[8] The inclusion of such a subjective condition into the definition of generation has been severely criticized by Castaneda as having no place in such an analysis [14].

[9] A reviewer correctly pointed out that this does not address how such causal connections are acquired nor how alternative equal explanations for the same event should be handled. Ultimately, this is an important capability for reasoning systems.

If we wish to handle branching acts and all of Goldman's examples, it appears we need something like (1)–(6) as the definition of generation. Picking the correct c, however, still remains problematic as one must examine *every* possible world. The counterfactual approach presented in this paper is more local and avoids the potential problems observed by Pollack. The definition for method-of seems to properly handle Goldman's examples as well as additional problematic cases. Firstly, branching acts and negative actions are handled properly through the explicit inclusion of clause (4). Secondly, since instances of clause (1) are separate from the definition of generation, no second order statements are needed and we therefore do not need (1), (5) or (6). Finally, the only purpose in ensuring the non-sufficiency of α in producing β ((2) and (2)), was to make that α and β were distinct, and this is made explicit in the new definition.

This treatment depends on a body of *basic generation knowledge* of the form: $\mathit{occurs}(\gamma, t) \wedge \mathit{holds}(c, t) \supset \mathit{occurs}(\delta, t)$. The inclusion of this basic knowledge does *not* mean that the counterfactual definition is superfluous—the reason is the same reason that 1–6 was not superfluous: generation, just like causation, suffers from the event subsumption problem. That is, the appropriate c is not always given *a priori*. Consider reports such as the following:

(16) By *leaving before the rush hour* he was able to arrive on time.

in which the italicized description is not one which is likely to be explicit in some element of the basic generation knowledge, just as in a causal analysis. Furthermore, the description *he was able to arrive on time* is again one which depends on the circumstances. A counterfactual analysis supplies a means for combining different types of knowledge—such as definitional knowledge for what it means to *leave before rush hour*—in a localized fashion to determine what the role, in this case, was of the event described by *leaving before rush hour*.[10] In short, the report (16) is true because the associated counterfactual, *if he had not left before rush hour, he would not have been able to arrive on time*, is true. In a similar vein, dealing with reports such as

(17) The flood was prevented by closing the gates.

would be difficult by reifying some causal primitive which specified all of the different circumstances in which a flood can be prevented: instead, these are computed as a function of the description of the means action. As another example, when reporting generation relations between negative actions or between actions whose description is a composite of the c's in the basic generation knowledge and other, possibly definitional knowledge, it is also difficult to pick the proper c. Again, the counterfactual approach is *localized* to the world and time in question and does not quantify over every possible world and time as the original Goldman-formulation does.

Notice that the definition given for method-of correctly handles the following example which is problematic for Goldman: *Bill jumped* $6'3''$. *John outjumped Bill by jumping* $7'$. Under a simple counterfactual analysis the best one case say is that John outjumped Bill

[10] This phenomenon exists whenever one modifies some event description with some temporal or spatial modifier.

by jumping over 6′3″; by incorporating the idea of strongest antecedent condition, this example is handled correctly.

6. Event prevention

This section extends a definition introduced elsewhere for the notion of event prevention [55]. The definition is based on a suggestion of Davis [23]: α prevents β just in case if α had not occurred, β would have been possible, while after α, β became impossible. The restriction to β being only possible and not necessary in the case of not(α) is there to handle cases such as:

(18) The vaccine prevented him from getting smallpox.

In this example, not receiving the vaccine does not necessarily mean that the referenced individual will acquire smallpox. The use of counterfactuals avoids problems with applying such a definition simply in terms of a branching temporal logic [50]. Cases involving events and not actions become problematic: if the future and past is fully determined by some set of causal laws then there is no branching temporal structure. Consequently, some counterfactual analysis is necessary (particularly in order to choose the "closest" world in which some event does *not* occur).

Definition 6.1 (*Event prevention*). An act-type, α, prevents an act-type, β, just in case,

$$\begin{aligned} &\mathit{occurs}(\alpha, t_1)\ \textbf{prevents}\ \mathit{occurs}(\beta, t_2) \equiv_{def} \\ &\quad \mathit{occurs}(\alpha, t_1) \rhd \neg \mathit{holds}(\Diamond \mathit{Occurs}(\beta, t_2), t_2) \end{aligned}$$

That is, if α occurs then, in all of the closest such worlds, β never occurs whereas if α does not occur then there is a possibility that β may occur. This assumes that the intended "context" is implicit. For example, a case such as

(19) I prevented him from speaking while he was in the room.

can be handled by either quantifying over the times in which the fact "he was in the room" is true or by describing the prevented action more fully. Other examples include situations in which an agent would not intend to perform some action: one would then assume that, under normal circumstances in that context, that action would not occur. For example,

(20) I prevented him from drinking this water by taking it away.

The notion of prevention here is much more restrictive in the sense that the prevented action is still possible. However, under the assumed context, it is not. A full treatment of context is beyond the scope of this paper [49,63].

It will be useful to define a prevents event type:

$$\begin{aligned} &\mathit{occurs}(\mathit{prevents}(\beta), t) \equiv_{def} \\ &\quad \exists \alpha \exists t' \geqslant t . \mathit{occurs}(\alpha, t) \wedge (\mathit{occurs}(\alpha, t)\ \textbf{prevents}\ \mathit{occurs}(\beta, t')) \end{aligned}$$

An important characteristic of Definition 6.1 is that it is not restricted to "positive" actions. That is, the two counterfactual clauses allow the definition to cover cases such as:

(21) Not opening the gate prevented the cattle from escaping.

since the second counterfactual clause in the definition requires that if the gate *had* been opened, the cattle would have escaped.

There are also situations which evoke their own "action" context: consider a traffic scenario and the statement [55]:

(22) The red light prevents him from turning left.

This statement follows from the restrictions imposed on the agent by the vehicle code.

7. Enablement

There are a number of ways in which one could define the notion of some action enabling some other action. One might consider the following option: α enables β just in case α brings about some condition, C, that is necessary for the performance of β. However, often there are several ways in which an action can be enabled. For example, suppose one has a bucket that one wishes to fill with water and there are two ways of filling the bucket, either by bringing the bucket to a nearby faucet or fetching a hose and using that to fill the bucket. Clearly, either of having the bucket under the faucet or having the hose near the bucket will enable the filling action. Suggesting that C represent a sufficient condition is also too strong: often there are several conditions that must jointly hold before some action can be performed. The definition will instead be framed around notions of possibility. [11] However, reports such as the following impose additional problems:

(23) Not stopping on the way home enabled him to arrive on time.

where the enabled action is already possible.

Definition 7.1 (*Enablement*).

$$\begin{aligned} &\mathit{occurs}(\alpha, t_1)\ \textbf{enables}\ \mathit{occurs}(\beta, t_2) \equiv_{def} \\ &\quad (\alpha \neq \beta \vee t_1 \neq t_2) \\ &\quad \wedge \mathit{occurs}(\alpha, t_1) \rhd \mathit{holds}(\Diamond \mathit{Occurs}(\beta, t_2), t_2) \end{aligned}$$

The use of counterfactuals ensures that if α had not occurred then β would not have been possible and it, therefore, rules out those circumstances in which β might have eventuated on its own. Example (23) can be handled by noting that there is no restriction that β be currently impossible. Notice that the truth of the counterfactual, *if he had stopped he would not have been able to arrive on time*, depends on restricting the set of possible actions to some action context as well as assuming everything else remained the same, the usual

[11] See Balkanski's [7] work for an approach based on generation conditions.

strategy adopted in the evaluation of counterfactuals. This would preclude the possibility of identifying aberrant events with the nonoccurrence of event α in the evaluation of example (23) so that when evaluating the second counterfactual, one does not consider quicker means of transport that would represent a departure from the norm.

Shoham [62] noted an equivalence between prevention and enablement: α enabled β just in case $not(\alpha)$ would have prevented β. This equivalence clearly holds among the definitions given here.

The above might be referred to as *event enablement* since one points to an event that causes the condition that enables; however, there is no reason not to point to the condition just as well. A report that points to an enabling event or an enabling condition seems to differ from a causing event or condition only in the sense that an enabling condition is best seen as part of the background: for example, in a situation involving the lighting of a match, we can say that the presence of oxygen enables the lighting of the match simply because that condition is part of some assumed background knowledge. In another situation, say a situation involving a vacuum in a laboratory where oxygen was accidently allowed to enter, the report: the oxygen caused the fire would be perfectly reasonable [16,48].

The distinction between foreground and background causes with respect to *actions* seems to have something to do with the intentionality of the enabled or caused action. It seems quite awkward to use *enables* when the enabled event was unintentional or viewed as undesirable by an external observer. Contrast:

(24a) Not picking up a ticket *caused* him to miss the train.

(24b) ?Not picking up a ticket *enabled* him to miss the train.

These two examples suggest that "causes" is a more appropriate term when describing an accident as the effect. However, other cases suggest that this distinction is not so clear cut:

(25) By flipping the switch he accidentally enabled the alarm.

It appears, then that some more complex scheme—at a pragmatic level of explanation—is necessary to fully distinguish foreground causes and background enablements.

8. Roles in group activities: Helping and hindering

Formalizing the notion of one agent "helping" another in some task illustrates the utility of counterfactual reasoning in isolating the role that an action plays in some particular situation. The intuition is that "helping" differs from "causing to" or "doing" in that it represents only one part of some cause. Before attempting to formalize this relation, let us consider a number of examples:

(26a) John helped him pick up the sofa. (As a result, he only needed one hand instead of two. In another scenario, he might not have been able to pick it up at all by himself.)

(26b) You can tell by the fact that it's a little easier if we turn the whole thing upside down [7]. (For example, when upside down, screwing on the plate does not require one to hold it in place.)

(26c) Smoothing the earth helped the logs roll down the slope [67].
(The logs would have instead taken more time to roll down the slope.)

(26d) Removing the benches helped the marchers cross the plaza [67].
(The marchers would have instead had to walk around the benches and required more steps to cross.)

In some cases the action responsible for helping occurs concurrently with the other action (e.g., (26a)) while in other cases the helping action is performed before (e.g., (26c) and (26d)).

These observations lead to the following definitions.

Definition 8.1 (*Helping*).

$$\begin{aligned} &occurs(\alpha, t_1)\ \textbf{help}\ occurs(\beta, t_2) \equiv_{def} \\ &\quad \exists r.in(resources(r), \beta) \\ &\quad \wedge occurs(\alpha, t_1)\ \textbf{enables}\ occurs(\beta, t_2) \end{aligned}$$

That is, some action, α, on the part of one agent helps in the performance of some other action, β, performed by another agent just in case: α enabled the performance of β with resources r; that is, β was of the form $e@resources(r)$. Notice that the agents of α and β might be identical. Neither is the possibility of those agents each representing a group precluded. [12]

This axiom assumes that what counts as a resource for an action is given. Resources are properties such as time taken, energy expended, steps taken, number of agents required, money, etc. [13] In this definition, the intent is that "not performing β with resources R will be treated as a sort of implicature [44] so that it will be taken to mean that the action involved more resources than those in R". This requires the following ancillary definitions.

Definition 8.2 (*Narrow negation*).

$$\begin{aligned} &\models occurs(not(\alpha@mods), t) \equiv \\ &\quad holds(name(\alpha) \\ &\quad \wedge \exists m \exists n.most\_specific(mods, m, n), t) \\ &\quad \wedge occurs(\alpha@n@not(m), t) \end{aligned}$$

where $most\_specific(mods, m, n)$ is true just in case n stands for all of the modifiers in *mods* minus the most specific one, m. The above says that one *not-*α*@mods*'s just in case

[12] See Fig. A.1 of Appendix A for the definition of the predicate "*in*".

[13] We could choose to instead define such a concept, perhaps as any object or property which is quantifiable and which, in addition, must necessarily be maintained under some threshold in order for the target action to be performed as described. Instead, this paper will assume the existence of knowledge that connects resources with actions. This is much the same as in decision theory which assumes the existence of a utility function. A full theory of resources that explained exactly how one decides on what counts as a resource is beyond the scope of this paper.

one α's in exactly the same way with the exception that the most specific description of α does not hold. For example, "not running home quickly" means one "runs home" but not quickly. This assumes that there is some way to scale action modifiers in terms of their specificity.

For the purposes of the formalization of *helping* we need only stipulate that the modifier, $resources(r)$, to an action, α—that is, to cases of $occurs(\alpha@agt(i)@mods@resources(r), t)$—is always assumed to be most specific.

The treatment of negation with respect to resources which are inherently quantities is treated as an implicature as follows.

Definition 8.3 (*Quantity implicature*).

$$\models occurs(\alpha@not(resources(r)), t) \equiv$$
$$\exists r'.occurs(\alpha@resources(r'), t) \wedge holds(r' > r, t)$$

That is, when we say that α-ing with r resources would not have been possible we mean that it would have required more than r resources to α.

Just as in the definition of prevention, relativizing the definition for helping to some context is useful in restricting the number of "irrelevant" possibilities that are considered. For example, in (26d) above, the context might constrain the methods by which one might consider the marchers "moving across the plaza" to not include unusual methods such as by helicopter. Additionally, the following point on colloquial usage of the term "help" should be made. When we say, as in (26a), that "John helped him pick up the sofa" what we really mean is that John helped or contributed to the joint action of picking up the sofa: in the end, John did not pick up the sofa, they both did.

The definition correctly handles the inherent opacity of such reports: if I helped someone move the piano quietly I might not actually have done any of the moving; I may have just kept the strings dampened. The former inference is blocked by virtue of the embedded counterfactual within the definition of enablement.

Hindering by way of some β is related to helping in the following way. When an action, β, is hindered, its performance entails a greater number of resources or cost. Examples of hindering include:

(27a) Mounds of earth hindered the logs in rolling down the slope [67].

(27b) The benches hindered the marchers in crossing the plaza (ibid).

One can define hindering straightforwardly in terms of of the definitions given so far.

Definition 8.4 (*Hindering*).

$$occurs(\alpha, t_1) \textbf{ hinders } occurs(\beta, t_2) \equiv_{def}$$
$$occurs(\alpha, t_1) \textbf{ help } occurs(not(\beta), t_2)$$

That is, some β is hindered just in case some α helps in increasing the number of resources used.

9. Resistance to action: Coercions

The case of a coerced or forced action is one characterized by some element of "resistance" on the part of the coerced agent or object. For example,

(28) He forced the unit into the casing with a screwdriver.

In this case, some abnormal amount of force/instrument is necessary in order to perform the indicated action. In the case of an agent being forced to perform some action, resistance can take the form of some desire or strong reason *not* to perform the target action. As such, coercions represent deviations from the perspective of the normal evolution of deliberations from desire to intention formation: they are sometimes characterized by an initial intention *not* to perform some action; this intention might subsequently be dropped because of the actions of some other agent. For example, some communicative action on the part of another agent might cause the agent to re-consider the reasons for forming the original intention and adopt the complementary intention. However, a desire not to perform the indicated action often remains. Consider:

(29a) Through his threats, the assailant forced me to open the safe.

(29b) The police forced everyone out of the park.

There are also examples of coercions in which the desire not to perform the action in question does not survive and in which the resistance to action takes on other forms:

(30) He forced John to change his mind about the importance of that tax credit.

Here again, there appears to be some departure from the norm having to do with the reasons the agent formed for his action. In (30), John's mental state might be such that he would normally maintain my belief regarding the importance of the tax credit, however, perhaps through some argument he became convinced otherwise.

Definition 9.1 (*Coercions*).

$$
\begin{aligned}
&\mathit{occurs}(\alpha, t_1)\ \textbf{forces}\ \mathit{occurs}(\beta, t_2) \equiv_{def} \\
&\quad \mathit{occurs}(\alpha, t_1)\ \textbf{causes}\ \mathit{occurs}(\beta, t_2) \\
&\quad \wedge \exists x \exists y . \beta = x@\mathit{resources}(r)@y \\
&\quad \wedge \mathit{holds}(\mathit{normal\_resources}(x@y, n), t)) \\
&\quad \wedge \exists \delta . \mathit{occurs}(\delta, t_1) \\
&\quad \wedge \mathit{occurs}(\delta, t_1)\ \textbf{hinders}\ \mathit{occurs}(x@\mathit{resources}(n)@y, t_2)
\end{aligned}
$$

This says that cases of coercions are cases in which a causal relation exists between the two target events and also there is some event, δ, which hinders the caused action from utilizing the number of resources that it would *normally* consume in that sort of situation: $\mathit{normal\_resources}(x@y, n)$.

Within this framework, the characteristic notion of "resistance" is captured in counterfactual terms through reference to resource consumption and not by way of some hypothetical force structure [67].

10. Non-movement and maintenance actions

Maintenance events differ from accomplishments in that they do not necessarily involve change with respect to the state of the object being maintained: by repeatedly pushing a door (call each such component event an α) one can maintain the door in a closed position (call this condition ϕ) [56].[14] The condition which is maintained, ϕ, must be counterfactually related to each component α: if one hadn't been pushing the door, it could have opened at some point, possibly because someone else was pushing from the other side. This cannot be determined simply by observing the scene: one might simply be pushing a locked door.

The properties discussed here can be captured by defining a maintenance of some condition, ϕ, as a process consisting of individual instances of hindering progress towards the achievement of $\neg\phi$. The following two axioms define what it means for an action to be identified as a "hindering" and also what it means for there to be no change relative to some ϕ.

$$\begin{aligned}
&\models \mathit{occurs}(\mathit{hindering@obj}(\phi)@\mathit{resources}(r)@\mathit{agt}(i), t) \equiv \\
&\quad \exists e.\mathit{occurs}(e@\mathit{agt}(i), t) \\
&\quad \wedge \mathit{occurs}(e@\mathit{agt}(i), t)\ \textbf{hinders}\ \mathit{occurs}(\mathit{progress@to}(\phi)@\mathit{resources}(r), t)
\end{aligned}$$

$$\begin{aligned}
&\models \mathit{occurs}(\mathit{wait@wrt}(\phi)@\mathit{agt}(i), t) \equiv \\
&\quad \exists x.\mathit{holds}(\mathit{basic}(i, x), t) \wedge \mathit{occurs}(x@\mathit{agt}(i); \phi?, t)
\end{aligned}$$

The act-type $\mathit{progress@to}(\psi)$ refers to the degree to which progress will be made towards the achievement of ψ: for example, if one is walking home then one's progress might be measured in terms of the number of steps along paths in the direction of home [57]. A hindering would then correspond to a situation in which the progress under some set of resources (steps taken) is increased.

A maintenance action can now be defined as follows.

Definition 10.1 (*Maintenance actions*). A maintenance event, $\mathit{maintain@obj}(\phi)$, is a process of the following form:

$$\begin{aligned}
&\models \mathit{occurs}(\mathit{maintain@agt}(i)@\mathit{obj}(\phi)@m, t) \equiv \\
&\quad \mathit{occurs}(\phi?; [\mathit{hindering@agt}(i)@\mathit{obj}(\neg\phi)@m \cup \mathit{wait@wrt}(\phi)@\mathit{agt}(i)]^*; \\
&\qquad\qquad [\mathit{hindering@agt}(i)@\mathit{obj}(\neg\phi)@m]; \\
&\qquad\qquad [\mathit{hindering@agt}(i)@\mathit{obj}(\neg\phi)@m \cup \mathit{wait@wrt}(\phi)@\mathit{agt}(i)]^*, t)
\end{aligned}$$

[14] This is also sometimes referred to as *protecting* a fact [50].

This definition identifies a maintenance action with a possibly inhomogeneous process consisting of instances of hindering the achievement of $\neg\phi$ with the further restriction that ϕ be true throughout. The second clause requires the occurrence of a hindering while the first and third clauses allow for the possibility that the referenced hindering is embedded in a process that might or might now involve additional hinderings (that is, an inhomogeneous process).

One particular virtue of the definition is that it explains reports involving the maintenance of some condition by way of *inaction*. For example, suppose a prisoner is being kept in an unlocked room for later questioning by police. Suppose further that a guard is stationed outside the door to the room. The following report

(31) The guard kept the prisoner in the room by not leaving the area.

is satisfactorily explained by Definition 10.1: if the guard *had* left the area then the prisoner might have escaped.

11. Rational agency: Causal connections between mind and action

Much of the impetus for belief, desire, and intention (BDI) models of rationality [15] has sprung from a certain brand of folk-psychological wisdom: rendering behavior intelligible amounts to identifying the reasons for which people act. This wisdom carries over as well to the determination of whether a particular description for an action is appropriate. In the sections which follow an approach to the characterization of rational action will be presented in which actions are grounded causally in a person's mental state at the time of action [16] [13,20]. Consider the simple report:

(32) He intentionally broke the vase.

Under the present suggestion, one could ascent to such a report just as long as the act of breaking the vase was caused by some intention; in its absence, the vase would have remained whole. Such an action could then be distinguished from an accidental breaking of the vase by noting the absence of an efficacious intention.

Such an approach requires that one first have a means for representing BDIs and then also have a way of attributing causal connections between mental states and action. The representation of beliefs and intentions will be examined briefly in this section; the former will be restricted to non-embedded belief reports. This will lead to a causal analysis of rational action, using the tools already developed in previous sections of this paper, to build a taxonomy of rational act-types.

An intention in $\mathcal{HL}$ can be represented as a fluent that ranges over agents and act-types: the $\mathcal{HL}$ statement $holds(Int(i,\alpha),t)$ is intended to mean that at time t agent i intends to α.

[15] For an overview of BDI theories see the volume edited by O'Hare and Jennings [54].

[16] Many of the important questions in this area were originally raised by philosophers [2,13,21,33]; they have since become of interest to many in AI as well [18]. The suggestion that the relation between mental states and action is causal has, in fact, been disputed by some philosophers [2].

To capture desirable properties of intentions and also to relate intentions to beliefs one can then build an $\mathcal{HL}$ theory making such properties and connections explicit. It is not possible to do justice in this paper to the immense body of work on the subject of intentionality; nevertheless, there are a number of important technical problems that must be addressed: chief among these is the *side effects problem*. It arises in connection with the following proposed role of intentions: intentions persist until they are achieved or believed impossible to achieve. The properties of persistence and commitment can be captured through EUT's treatment in $\mathcal{HL}$ of the frame problem. The *side-effects* problem is then, roughly, that one does not necessarily intend all of the side-effects of one's intentions [17]. An oft-quoted scenario meant to illustrate this problem is the following involving a patient who intends to go to the dentist, aware that it will be a painful visit. If the agent's side-effects are also intended and persist until satisfied then if the agent goes to the dentist and does not receive pain the agent will seek it out in some other way in order to satisfy the additional intention! This difficulty does not arise in the present setting because intention is taken as a primitive; in $\mathcal{HL}$ action types are identified with the set of world-interval pairs in which the action type occurs: even if the occurrence of β always follows the occurrence of α, where the former is intended, β will not necessarily also be intended unless the implication is bi-directional.

Bratman [13] argues that beliefs should be connected to intentions in the following way: an agent should not simultaneously intend to α while believing that α is not possible. This requirement on the agent's rationality can be captured by the following.

Rationality Postulate 11.1 (*Intentions and beliefs*). If an agent, i, intends some α then the agent will not believe that α will never be possible.

$$\models \mathit{holds}(\mathit{Int}(i,\alpha),t) \supset \exists t' \geqslant t.\mathit{holds}(\Diamond_i \mathit{Occurs}(\alpha,t'),t)$$

Intentions should also, according to Bratman, be internally consistent. That is, if one intends α then one cannot also intend to $\mathit{not}(\alpha)$. This is captured in the following axiom.

Definition 11.1 (*Consistency of intentions*).

$$\models \mathit{holds}(\mathit{Int}(i,\alpha),t) \supset \neg \mathit{holds}(\mathit{Int}(i,\mathit{not}(\alpha)),t)$$
$$\forall \alpha.\mathit{not}(\mathit{not}(\alpha)) = \alpha$$

Definition 11.2 (*Dropping an intention*).

$$\mathit{holds}(\mathit{Int}(i,\alpha),t) \wedge (\mathit{occurs}(\alpha@\mathit{agt}(i),t) \vee \forall s \geqslant t.\neg \mathit{holds}(\Diamond_i \mathit{Occurs}(\alpha,s),t)) \supset \neg \mathit{holds}(\mathit{Int}(i,\alpha),t+1)$$

That is, an agent will drop an intention as soon as it performs the action or believes it is not possible.

We want to connect action to mind. The first step is to stipulate that all actions are grounded in some basic action. For example, one might move a chess piece to a certain square by moving one's arm in a certain way (the latter representing a basic action). In this way, a bona-fide agentive action (such as moving the chess piece) can be distinguished from a spurious one (such as one in which someone takes one's hand and forces it to make the same motions).

Definition 11.3 (*Agentive actions*).

$$\begin{aligned}&\models \mathit{occurs}(\alpha@\mathit{agt}(i), t) \equiv\\ &\quad \mathit{holds}(\mathit{basic}(i, \alpha), t)\\ &\quad \vee (\exists\beta.\mathit{occurs}(\mathit{by}(\alpha, \beta), t) \wedge \mathit{holds}(\mathit{basic}(i, \beta), t))\end{aligned}$$

That is, an action is either basic or grounded in some basic action.

Basic actions are further viewed here as intended [13].

Definition 11.4 (*Basic actions are intended*).

$$\models \mathit{occurs}(\alpha@\mathit{agt}(i), t) \wedge \mathit{holds}(\mathit{basic}(i, \alpha), t) \supset \mathit{holds}(\mathit{Int}(i, \alpha), t)$$

As already mentioned, the connection between intention and intentional action pursued in this paper will be a causal one along the lines put forward by Bratman [13]: to perform α intentionally is to α while intending α or some "related" action. [17] However, many side effects of intentions, even though not intended, are nevertheless viewed as intentional. For example, consider an agent who intends to purchase some item with a credit card under the circumstances (call these, C) in which the purchase will put the agent at his credit limit. Then, as long as the agent is aware that C holds, he will have intentionally made the purchase and intentionally pushed his credit card to the limit even though the latter might not have been intended. In particular, if the agent discovered, while trying to purchase the item, that it was being given away, the agent would not have necessarily sought some other way to increase his credit card balance. However, had the agent not been aware that C held, he would not have been accused of intentionally performing the indicated side-effect. Another well-known example is the following involving someone who intends to kill a particular person. The killer, after having formed the intention to kill, sets out to drive to the person's home to carry out his intention. Along the way, he looses control of his car and strikes and kills a pedestrian on the sidewalk. To his surprise, the person he strikes turns out to be the person he had intended to kill. However, it would be strange to claim that the person had killed the pedestrian intentionally.

The prior intention must also be linked *causally* to the referenced action. Consider an example: an agent intends to find out his file system quota but doesn't know how. The agent also intends to find out what his current file usage is; he knows he can perform the latter in UNIX by issuing the command `df`. Before finding out how to satisfy the first

[17] This is not the only path possible, however; certain philosophers have equated intentional action with being in a particular relation to some belief-desire complex, where desires are, in a sense, weaker than intentions.

intention he proceeds to satisfy the second and, to his surprise, the same UNIX command supplies the information to satisfy both intentions. Did he thereby intentionally discover his quota? It seems here that what is missing is a causal connection between the intention and the action: the intention to discover his file system quota is not what prompted him to type `df`.

These cases can be handled as follows.

Definition 11.5 (*Intentional action*).

$$\begin{aligned}&\models \mathit{occurs}(\alpha@\mathit{agt}(i)@\mathit{manner}(\mathit{intentional}), t) \equiv\\ &\quad \exists\beta.\mathit{holds}(\mathit{Int}(i,\beta), t)\ \textbf{caused}\ \mathit{occurs}(\beta@\mathit{agt}(i), t)\\ &\quad \wedge \mathit{holds}(\mathit{Know}(i, \mathit{Occurs}(\beta, t)\ \textbf{Method}\ \ \mathit{Occurs}(\alpha, t)), t)\end{aligned}$$

where [18]

$$w \models \mathit{holds}(\mathit{Know}(i, \mathit{Occurs}(\beta, t)\ \textbf{Method}\ \mathit{Occurs}(\alpha, t)), t')$$

iff

$$\sigma_i(w, t') \models \mathit{occurs}(\beta, t)\ \textbf{method}\ \mathit{occurs}(\alpha, t)$$

and

$$w \models \mathit{occurs}(\beta, t)\ \textbf{method}\ \mathit{occurs}(\alpha, t)$$

That is, the agent's intention to β causes him to β (this solves the third objection) which, in turn, is known by the agent to represent a means for α (this solves the first two objections since only *believed* side-effects are performed intentionally). The second clause is based on a suggestion and analysis presented by Bratman [13].

Intentions have been observed to crop up in action reports in other ways [5,13]. These will not be examined here. One brand of report involves *acting with an intention*:

(33) I rented the tuxedo with the intention of returning it the next day.

In examples such as this, the action of "renting the tuxedo" is part of some larger plan which includes, among other things, the plan to return the tuxedo. Austin observes that such an action should be "judged" relative to such a plan [5]: if the agent does not return it the next day, some explanation is necessary. In this way, plans can serve as contexts for the interpretation of actions [8,31]. A second sort of case involves *acting for a purpose*:

(34) I rented the tuxedo for the purpose of wearing it at the wedding.

Here, the purpose ("wearing it at the wedding") would seem to guide the formation of further intentions such as that of intending to have a tuxedo for the wedding. In contrast, the "intention of returning it the next day" in the first example could not act as the purpose for renting the tuxedo.

Some additional useful axioms include:

[18] See also Appendix B.

$$\models holds(Int(i, \alpha; \beta), t) \supset$$
$$holds(Int(i, \alpha), t) \wedge holds(Int(i, \beta@after(\alpha)), t)$$

$$\models holds(Int(i, Int(i, \alpha)?), t) \supset holds(Int(i, \alpha), t)$$

$$\models holds(Int(i, by(\alpha, \beta)), t) \supset holds(Int(i, \alpha) \wedge Int(i, \beta), t) \quad (7)$$

12. Range of responsibility: Accidents and mistakes

Cases of accidents, failures,[19] mistakes, and coercions as *deviations from the normal causal pathway of rational action*. The term causal pathway will be used to refer to the sequences of changes in an agent's mental state that causes an agent to act in a certain way.

Let me begin with the case of an accident. Consider some typical examples.

(35a) He spilled the coffee when he picked it up.

(35b) He accidentally insulted John when he spoke to him.

(35c) He accidentally arrived late to the party because he took the wrong turn.

In the first example, the agent may or may not have been trying[20] to *not* spill the coffee. However, in either case the spilling of the coffee was unintentional. In (35b), there may have been some unknown fact about the circumstances—say, John's mental state and his concomitant dislike for the subject matter the other agent was about to engage him in—that resulted in John becoming insulted when spoken to. In the third case, the agent may have chosen some incorrect means action which resulted in getting lost. The means action was incorrect due, again, to some incompleteness in knowledge of the circumstances. In each of these cases, the accident was unintentional and furthermore the agent may or may not have been trying not to get lost, etc. In some cases, say (35b), it might be that if the agent had known that he was going to insult John, he would have tried not to. This is not always the case, however. Consider instances of "lucky" accidents.

(36) I accidentally bumped into him before the meeting, and I'm glad I did.

In this case, no regret is associated with the accidental eventuality. Another example of a lucky accident is that given in Section 11 involving someone who succeeds in killing someone but not in the manner intended. In that case, if the agent knew he was going to kill in the particular manner in which he did he might have done so anyway.

Definition 12.1 (*Accidents*).

$$\models occurs(\alpha@agt(i)@manner(accident), t) \equiv$$
$$occurs(\alpha@agt(i), t) \wedge \neg occurs(\alpha@agt(i)@manner(intentional), t)$$

[19] To be discussed in the next section.

[20] Cases of attempts are analyzed later in relation to failures. Briefly, an attempt to α is characterized by the performance of some β together with the intention to α by β.

In the case of non-basic α, the above, together with the definition of an intentional action, says that an agent, i, accidentally α's just in case either:

(i) some β that represented a method for α was not caused by an intention to β, or
(ii) the agent did not know that β represented a method for α.

In this definition, the agent may or many not have been trying to not-α. The former case will be referred to as a failure. These are discussed shortly. An example of an accident which is not a failure is example (35b) in which the agent need not have had any intention to not insult John. Cases of "unlucky" accidents play an important role in the characterization of mistakes. A regrettable or unlucky accident is one which an agent would have otherwise "taken back".

Definition 12.2 (*Regrettable accidents*).

$$\begin{aligned}
&\models \mathit{occurs}(\alpha@\mathit{agt}(i)@\mathit{manner}(\mathit{accident})@\mathit{modifier}(\mathit{regrettable}), t) \equiv \\
&\quad \mathit{occurs}(\alpha@\mathit{agt}(i)@\mathit{manner}@\mathit{accident}, t) \\
&\quad \wedge \exists\beta.\mathit{occurs}(\mathit{by}(\alpha, \beta)@\mathit{agt}(i), t) \\
&\quad \wedge \neg\mathit{holds}(\mathit{Know}(i, \mathit{Occurs}(\beta@\mathit{agt}(i), t)\ \textbf{Method}\ \mathit{Occurs}(\alpha@\mathit{agt}(i), t)), t) \\
&\qquad \textbf{caused}\ \mathit{occurs}(\beta@\mathit{agt}(i), t)
\end{aligned}$$

Interestingly, this definition illustrates a case in which an agent's action is *caused* by an absence of a particular attitude.

Whereas an accident is a characterization of an ends action, a *mistake* is a characterization of a means action.[21] In (35a), holding the cup in a certain way may have been the crucial mistake execution that resulted in the accidental spilling. In (35b), the agent speaking to John in the way that he did was a mistake and resulted in accidentally insulting him. Finally, in (35c), taking the particular turn on the way to the party was a mistake: it resulted in accidentally arriving late. In each case, the choice of action was incorrect. We therefore have,

Definition 12.3 (*Mistakes*).

$$\begin{aligned}
&\models \mathit{occurs}(\alpha@\mathit{agt}(i)@\mathit{manner}(\mathit{mistake}), t) \equiv \\
&\quad \exists\beta.\mathit{occurs}(\beta@\mathit{agt}(i)@\mathit{manner}(\mathit{accident})@\mathit{manner}(\mathit{regrettable}), t) \\
&\quad \wedge\ \mathit{occurs}(\mathit{by}(\beta, \alpha)@\mathit{agt}(i), t) \\
&\quad \wedge\ \exists\gamma.\mathit{holds}(\Diamond_{\Omega}\mathit{Occurs}(\mathit{by}(\mathit{not}(\beta)@\mathit{agt}(i), \gamma), t), t)
\end{aligned}$$

That is, an agent mistakingly α's whenever it accidentally and regrettably β-ed by α-ing. The last clause requires as well that it be physically possible to *not*-β.

For an interesting philosophical discussion of such act-types, the interested reader is directed to Austin's work [4,5].

[21] This was observed in the implementation described later.

13. Attempts and failures

Whenever we ascribe an instance of *trying-to-*α to some agent, i, we seem to suggest that i performed some β which it believed would result in α. If i fails then either:

(i) some expected circumstance necessary for β to represent a method for α did not obtain,

(ii) i's beliefs about the circumstances in which it was embedded were correct but its beliefs about the relation between β and α were incorrect, or

(iii) i failed to perform β correctly. [22]

For example, consider the following:

(37a) John tried to escape but was caught.

(37b) John tried to remove the stains with soap and water.

(37c) John tried not to spill the coffee by holding the cup steady but failed.

In the first example, we can imagine a situation in which John attempts an escape by executing some plan of action, β, believing that he will thereby escape. However, the circumstances might be such that β cannot generate the desired action: suppose, for example, that unbeknownst to John someone is positioned in such a way as to prevent the escape; in this case, John's inaccurate beliefs about the world prevent him from accurately predicting the future. In (37b), John might have perfect knowledge regarding the current situation but his beliefs concerning possible means for removing stains could be incorrect. Finally, in the last example, John's beliefs about the relation of holding the cup steady and preventing the spilling of the coffee are correct, as are his beliefs about the current situation; in this case, however, he simply fails to perform the action *hold cup steady* properly.

The following simple axiom captures these observations:

Definition 13.1 (*Attempts*).

$$\models \mathit{occurs}(\mathit{try}(\alpha)@\mathit{agt}(i), t) \equiv$$
$$\exists\beta.\mathit{holds}(\mathit{Int}(i, \mathit{by}(\alpha, \beta)@\mathit{agt}(i)), t) \textbf{ caused } \mathit{occurs}(\beta@\mathit{agt}(i), t)$$

That is, an agent i can be said to try to α just in case i performs some β with the intention of performing α. The agent need not have a firm belief that by β-ing it will succeed in α-ing: the agent must only not believe that it won't succeed. Under this definition, examples such as (37c) would be analyzed as a case of $\mathit{try}(\mathit{by}(\mathit{not}(\mathit{spill}), \mathit{steady}))$ where the agent performed some γ it incorrectly believed would represent a method for the hold-steady action. Once again, the inclusion of the *caused* clause ensures that the agent was not forced to β by, for example, someone moving the agent's hand.

There is a potential problem with this definition, however. This stems from an observation of Austin's [3]. If one adopts a notion of ability along the lines of Moore [52] then for an agent to be *able* to α, the agent must know of some basic (executable) action that denotes β such that β will result in α. However, such a formulation together with the above

[22] Pollack [59] discusses related issues in the context of plan recognition.

analysis of attempts and failures results in agents that are infallible with respect to those α that fall within the domain of their capabilities. As Austin notes, one might be justified in claiming that one can make a particularly short golf putt without it, however, also being the case that one necessarily succeeds. Perhaps the distinction necessary is that of *know-how-to-perform* and *can-perform* [53, Chapter 6], so that the agent isn't guaranteed success just because he knows an executable specification of the action; the physical preconditions have to be satisfied as well, along with a set of protocols (including social protocols that govern how rational and civilized agents behave).

The notion of a failure can now be captured as follows:

Definition 13.2 (*Failures*).

$$\models occurs(fail(\alpha)@agt(i), t) \equiv$$
$$occurs(try(\alpha)@agt(i), t) \wedge \exists\beta.occurs(by(not(\alpha), \beta)@agt(i), t)$$

where

$$\models occurs(fail(\alpha), t) \supset occurs(not(\alpha), t)$$

"You can't fail if you've never tried", as they say. Notice that the physical possibility of α is not necessary for a failure, all that is needed is simply a deviation from expectations.

The theorem that follows requires the following conjecture that an agent's actions are tied to his beliefs and intentions. It assumes that the agent has already performed the means-end reasoning to arrive at a plan for action.

Definition 13.3 (*Causal pathway of rational action*). This postulate says that if an agent has a plan to perform some action and believes that it can succeed, then that plan will normally be translated into action.

$$\models holds(Int(i, by(\alpha, \beta)), t) \wedge holds(basic\_seq(i, \beta), t)$$
$$\wedge holds(\Diamond_i Occurs(by(\alpha, \beta)@agt(i), t), t) \wedge \neg holds(ab(\beta@agt(i)), t)$$
$$\supset occurs(\beta@agt(i), t)$$

The $ab(\cdot)$ term is introduced as a way of rendering the rule defeasible [58]. This assumes that intentions are already decomposed into basic acts, where

$$holds(basic\_seq(i, \gamma), t) \equiv$$
$$holds(basic(i, \gamma) \vee (\gamma = \alpha; \beta \wedge basic(i, \alpha) \wedge basic\_seq(i, \beta)), t)$$

and

$$holds(ab(\alpha; \beta) \equiv ab(\alpha) \wedge ab(\beta), t)$$
$$holds(basic(i, \alpha), t) \equiv holds(Bel(i, basic(i, \alpha)), t)$$

A consequence of the definition for a failure is the following.

Theorem 13.1 (*Relation between Failures and Accidents*).

$$\models occurs(fail(\alpha)@agt(i), t) \supset occurs(not(\alpha)@manner(accident)@agt(i), t)$$

That is, if an agent fail(α)-ed then, by the definition, the agent failed to α when it was trying to α. This theorem says that the agent must also have accidentally not(α)-ed. That is, all failures are accidents but not necessarily vice versa.

Proof. By definition of a failure, i intended to α by some β but instead not-α-ed.

Claim. *i not-α-ed unintentionally.*

Suppose instead that i not-α-ed intentionally. Then, by the definition of an intentional action the agent must have intended some γ which caused it to γ and, moreover, i was aware that γ represented a method for not-α. Therefore, in all $w' \in \sigma$, $w' \models$ $occurs(\gamma, t)$ **method** $occurs(not(\alpha), t) \wedge occurs(not(\alpha), t)$. But since an agent's intentions must be consistent, $\neg holds(\Diamond Occurs(\alpha, t), t)$, which is a contradiction. □

One consequence of the theory so far is that failures are counterfactually grounded in either some more basic action or some partial mental state description. This provides a uniform treatment of *how* questions: to explain how someone failed to α, one either refers to a more basic action to which the failure is counterfactually related (i.e., "he failed to α by γ-ing, when he should have β-ed") or to either the agent's lack of know-how, lack of situational information, or lack of know-how with respect to some more basic action. Note also that agents can be successful while at the same time lucky: a belief that some β represents a method for α might be true but unjustified.

Theorem 13.2 (Agents will never try to fail).

$$\models \forall t \forall \alpha. \neg occurs(try(fail(\alpha)), t)$$

Proof. Suppose not. Then there is some e and s such that $occurs(try(fail(e)), s)$. By the definition of an attempt we have:

$$holds(Int(i, by(fail(e), \beta)), s) \wedge occurs(\beta, s)$$

for some β. By the closure property of intentions (7) with respect to actions, we then have:

$$holds(Int(i, fail(e)), s)$$

Now, by the definition of a failure and the definition of a sequence and test action we have that this implies:

$$holds(Int(i, (Int(i, e)?; not(e))), s)$$

$$holds(Int(i, (Int(i, e)?)), s) \wedge holds(Int(i, not(e)@after(Int(i, e)?)), s)$$

But since the duration of a test action is zero and since an intention to intend e is the same as intending to e, this implies:

$$holds(Int(i, e), s) \wedge holds(Int(i, not(e)), s)$$

This is a contradiction. □

Finally, there is an idiomatic use of *fails* in English which falls more properly under the rubric of a neglecting to perform some action when one should have, rather than failure. Consider the following:

(38) John failed to pick up the beer for the party.

in which we cannot identify any other particular action on the part of John which he performed with the intention of picking up the beer for the party.

14. Method-of relations between negative actions and cases of *letting*

The semantics of the method-of relation has been given in terms of counterfactuals; the use of counterfactuals was seen as an explanatory tool. Here, cases in which one would be justified in saying that some agent not-α-ed by β-ing, where β can represent a "positive" or "negative" action are considered in more detail. At the same time, cases which suggest a means for fixing the time of a negative action will be examined.

When a negative description, $not(\alpha)$, appears in the second argument of a causal report, the relationship represented is often a prevention. As an example, consider a scenario in which someone is trying to catch a plane and stopping to telephone the gate provided the only means of achieving that goal. Here, we have:

(39a) Not stopping to telephone prevented the passenger from arriving at the gate on time.

(39b) Not stopping to telephone caused the passenger to arrive late.

Once again, these statements can be evaluated by way of counterfactuals. In both cases, one could ascribe a negative action to the agent in question: *his not arriving on time* which was caused by his *not stopping to telephone*. However, not all preventions can be identified with a negative action. Cases which are generally are those such as (39) and (40) in which a negative action comes about as a side-effect *during* the performance of some other action.

(40) He avoided spilling the coffee by holding the saucer steady.

As discussed in the analysis culminating in Axiom 10.1, we have here that the agent in question is involved in some process (say, transporting the cup from one location to another) by which it brings about some ϕ. Interleaved with this process is some secondary process, say e, which prevents the bringing about of the spilling. Examples (39) and (40) both suggest that α is a method for not-β just in case not-α and β are in counterfactual dependence. In these sorts of cases the time of the not-β action is coextensive with the time of α. Similarly, in cases involving some negative action that represents a method for some positive action, such as:

(41) By not lowering his hand he signaled again to the auctioneer.

we might fix the time of the not-lowering action to the interval over which the signal occurred (presumably corresponding to the span over which the auctioneer observed the raised hand or the period between the last bid and the current one).

Cases of one agent letting another agent perform some action represent a special type of negative action: there is a strong intuition that when an agent let's something happen he refrains from performing some other action. It is important, therefore, to consider examples of such events. Let us examine the following candidate definition for such an act-type which, although it has problems, is pretty close to what we want:

$$\models occurs(let(\beta)@agt(i), t) \equiv$$
$$occurs(not(prevents(\beta))@manner(refrain)@agt(i), t)$$

where a *refrain* is defined as

Definition 14.1 (*Refrains*).

$$\models occurs(not(\alpha)@manner(refrain)@agt(i), t) \equiv$$
$$\neg occurs(\alpha@agt(i), t) \wedge holds(\Diamond_i Occurs(\alpha@agt(i), t), t)$$

That is, agent i refrains from α just in case α is possible but the agent did not perform it. [23]

Therefore, we have that a letting is just refraining from a prevention when that prevention is possible. The action not prevented, α, must also, by virtue of the definition of prevention, be possible though not necessarily necessary. This is consistent with the observation made by Jackendoff in [38] (his Example (28), p. 135):

(42a) Harry let Sam leave, and so Sam left.

(42b) Harry let Sam leave, but for some strange reason, Sam didn't leave.

since if an action is not prevented it is only possible but not necessary. It is crucial in the above provisional definition for letting that the additional restriction, *manner*(*refrain*), not be left out; otherwise, the prevention might not have occurred simply because the action was not possible in the first place.

However, there is a problem with the above definition for letting. It is too weak in the sense that it fails to distinguish cases of letting α from cases in which an agent causes α. This is because if causing α is true then indeed not preventing α is true also. To handle this, the above definition can be modified so that it is not dependent on the existence of the responsible agent.

Definition 14.2 (*Acts of letting*).

$$\models occurs(let(\beta)@agt(i)), t) \equiv$$
$$\wedge \exists\alpha.occurs(by(not(prevents(\beta))@manner(refrain), \alpha)@agt(i), t)$$
$$\wedge \neg(occurs(\alpha, t) \textbf{ instrumental } occurs(\beta, t))$$

[23] There is a sense in which a statement which reports that, for example, *John refrained from going to class today*, carries with it the expectation that the situation in question represented some sort of deviation from the norm; that is, that under normal circumstances the agent *would* have attended class; also perhaps there is the suggestion that John resisted going to class. This paper is not concerned with the precise meanings of verbs that report such instances of a negative action, but rather with a claimed common characteristic that they share: the expression of a counterfactual dependence.

That is, agent i must have refrained from preventing β, when a prevention was possible, and also if β did indeed occur, in those closest counterfactual worlds in which the agent does not exist, β occurs also. This latter clause ensures that agent i had nothing to do with initiating α.

The first conjunct therefore correctly handles cases such as

(43) He let the tank overflow by not pulling the plug.

while the second correctly blocks cases such as

(44) #He let the tank overflow by filling it.

The second is problematic without the second clause because although the agent might have refrained from preventing the tank from overflowing (where the act of preventing would have precluded a number of possible futures resulting in the overflowing, such as by permanently shutting off the water supply) while also causing the tank to overflow by actually filling it. This is a good example of using counterfactuals to determine the *role* that an agent or action plays in some eventuality.

Certain cases require attention to the extent and composition of the events under analysis. Consider

(45a) He let the tank empty by pulling the plug.

(45b) He let the car slide off the cliff by kicking the rock (and, so, it rolled down, uninteruppted).

Both cases would be handled under the present theory as cases in which an action enabled another which marked the start of some process. In both cases, however, the agent was not instrumental in the continuation or culmination of of the process (that is, the crashing to the bottom in example (45b). Notice that constraining the definition to intentional lets to handle these last cases would be overly-restrictive as there are certainly cases of accidental lets.

15. Example

Suppose a certain prisoner is trying to escape from a room in which he is being detained. There are two exits: a front and a rear exit, both of which are locked. If there is a guard posted at an exit, the prisoner cannot escape. The prisoner is, however, clever enough to be able to pick the lock. The prisoner decides to take the rear exit: he unlocks the door and exits but, as it turns out, a guard is posted at that exit, though not at the front.

One way to formalize this scenario is as follows (in which, for simplicity, reference to the agent of an action and of an intention is left implicit). Let the actual world be w_0 and let the action of unlocking the door occur at time 1. Let w_0 stand for the initial world and the set of causal laws at time 1, $\mathcal{L}(w_0, 1)$, be:

$$
\begin{aligned}
&\mathcal{L}(w_0, 1) = \\
&\quad \{holds(locked(x), t) \wedge occurs(unlock@obj(x), t) \supset \neg holds(locked(x), t+1), \\
&\quad\ occurs(exit@obj(x), t) \supset \neg holds(locked(x), t), \\
&\quad\ occurs(unlock@obj(x); exit@obj(x), t) \wedge \neg holds(guard(x), t) \\
&\qquad \supset occurs(escape@dur(2), t), \\
&\quad\ \neg occurs(unlock@obj(x), t-1) \wedge occurs(exit@obj(x), t) \wedge \neg holds(guard(x), t) \\
&\qquad \supset occurs(escape@dur(1), t)\} \\
&\quad \cup \{x = exit \vee x = escape \vee x = Rear \vee x = Front\} \\
&\quad \cup \{x \neq y \mid x \in Fluents = \{Rear, Front, Exit, Escape\}\ \&\ y \in Fluents - \{x\}\} \\
&\quad \cup \{t \neq t' \mid t \in \mathbb{N}\ \&\ t' \in \mathbb{N} - \{t\}\} \\
&\quad \cup \{t = 1 \vee t = 2 \vee \cdots \vee t = n \mid n \in \mathbb{N}\} \cup \mathcal{R}
\end{aligned}
$$

The first axiom states the effects of unlocking a door while the second specifies the preconditions of an exit through a particular door: it must be unlocked. The next two axioms define the action of escaping: to occur there must not be a guard at the door. If the door is locked the escape takes one extra time unit. The final four axioms are unique names assumptions and domain closure axioms where $\mathbb{N}$ stands for the set of integers. The symbol $\mathcal{R}$ stands for the set consisting of the axioms given in this paper for defining attempts, failures, preventions, and so forth, together with appropriate axioms for addition.

The world description, from the prisoner's point of view, consists of the following statements (the prisoner is assumed to be aware of the locked state of the door). For simplicity, reference to the agent is suppressed.

$$
\begin{aligned}
&\mathcal{WD}(w_0, 1) = \\
&\quad \{\neg holds(ab(unlock@obj(Rear)), 1), \neg holds(ab(exit@obj(Rear)), 2), \\
&\quad\ holds(locked(Front), 1), holds(ab(exit@obj(Front)), 1), \\
&\quad\ holds(locked(Rear), 1), holds(basic(unlock@obj(x)), t), \\
&\quad\ holds(basic(exit@obj(y)), t), \\
&\quad\ holds(Int(by(escape, unlock@obj(Rear); exit@obj(Rear))), 1)\}
\end{aligned}
$$

No assumptions are made involving the presence of a guard: this reflects the prisoner's uncertainty. To distinguish the prisoner's beliefs from the truth, reference is made to Ω, standing for an omniscient agent:

$$
\begin{aligned}
&\mathcal{WD}_{\Omega}(w_0, 1) = \\
&\quad \mathcal{WD}(w_0, 1) \cup \{holds(guard(Rear), 1), \neg holds(guard(Front), 1)\}
\end{aligned}
$$

and $\mathcal{L}_{\Omega}(w_0, 1) = \mathcal{L}(w_0, 1)$. Causal judgments are made relative to $\mathcal{WD}_{\Omega}$. Notice that $occurs(unlock@obj(Rear); exit@obj(Rear), 1)$ is not included in $\mathcal{WD}_{\Omega}$ as it will follow from rationality postulate 13.3; this is necessary in syntactic approaches to belief change and counterfactual reasoning [28,58]: the idea is that if an agent's beliefs are represented by

the logical closure of $A = \{p, p \supset q\}$, then the withdrawal of belief in p is represented by first removing p from A and then taking the logical closure; q will then no longer follow.

Given this description, the following theorems are typical of the sorts of commonsense inferences one might like to draw. [24]

Theorem 15.1. *"He tried to escape but failed". That is,*

$$occurs(try(escape), 1) \wedge occurs(fail(escape), 1)$$

Proof. We need to show that,

$$w_0 \models holds(Int(by(escape, unlock@obj(Rear); exit@obj(Rear))), 1)$$
$$\textbf{causes}\ occurs(unlock@obj(Rear); exit@obj(Rear), 1)$$

By Definition 3.3 for **causes**, we need to show that:

$$\sigma_\Omega(w_0, 1) \models occurs(unlock@obj(Rear); exit@obj(Rear), 1)$$
$$\wedge\, holds(Int(by(escape, unlock@obj(Rear); exit@obj(Rear))), 1)$$
$$\wedge\, (\neg holds(Int(by(escape, unlock@obj(Rear); exit@obj(Rear))), 1)$$
$$> \neg occurs(unlock@obj(Rear); exit@obj(Rear), 1)) \tag{8}$$

Computing the information state, $\sigma_\Omega(w_0, 1)$, is straightforward for this example [58]. Let $\sigma_\Omega(w_0, 1) = \{w'\}$, where the set of assumptions associated with w' is $\Gamma(w') = \mathcal{L}_\Omega(w_0, 1) \cup \mathcal{WD}_\Omega(w_0, 1) \cup \mathcal{P}_\Omega(w_0, 1)$, such that:

$$\mathcal{P}_\Omega(w_0, 1) = \{holds(\phi, t) \supset holds(\phi, t) \mid \forall t \forall \phi \text{ such that, } \phi \neq Locked, t \neq 1\}$$

The $\mathcal{P}_\Omega(w_0, 1)$ are persistence assumptions for handling the frame problem. They can be computed [58] or simply assumed for this example. The first conjunct in (8) then follows directly and the second is given. By the definition of the $>$ connective, we then have that

$$\sigma_\Omega(w_0, 1) \models \neg holds(Int(by(escape, unlock@obj(Rear); exit@obj(Rear))), 1)$$
$$> \neg occurs(unlock@obj(Rear); exit@obj(Rear), 1)$$

iff

$$min\{\sigma_\Omega(w_0, 1) \diamond \| \neg holds(Int(by(escape, unlock@obj(Rear);$$
$$exit@obj(Rear))), 1) \|, \preceq_E\}$$
$$\models \neg occurs(unlock@obj(Rear); exit@obj(Rear), 1)$$

One can demonstrate the truth of the counterfactual as follows. Let *Ab* stand for the set of negated *ab*-statements. In this example, the $\sqsubseteq$-ordering will then give lowest preference to worlds in which *Ab* holds. In updating σ, the $\sqsubseteq$-ordering specifies that one keep as much of the Γ set of assumptions as consistent with the counterfactual supposition. This means that one will prefer those worlds in which the intention to unlock the door and exit is not included in the new Γ and, hence, the fact that the agent performed those two actions will not

[24] The appendix discusses the semantics of counterfactuals and information states (σ).

follow. One then minimizes the set of unsupported positive facts using the $\preceq_E$-ordering: since the unlocking and exiting actions do not have support, we prefer worlds in which those actions do *not* occur. Notice that without this additional requirement, both worlds in which that belief hold and worlds in which it does not would be consistent with the counterfactual supposition. To show failure, we need only notice that $occurs(not(escape), 1)$ follows as well from σ, demonstrating the existence of a failure. □

Theorem 15.2. "*The agent accidentally and regrettably failed to escape*". *That is,*

$$\sigma_\Omega(w_0, 1) \models occurs(not(escape)@manner(accident)@modifier(regrettable), 1)$$

Proof. By Definition 12.2, one first needs to show that the action was accidental: that is, that it was not intentional. We are given that the agent intended to escape; therefore, by Definition 11.1 and Axiom (7), the prisoner would not have intended to not escape. Regrettability requires that one first find some β that represented a method for not-escaping. Let β stand for $unlock@obj(Rear); exit@obj(Rear)$, or "unlocking and exiting the rear, guarded exit". By Definition 4.2 for method-of, we need to show that

$$\sigma_\Omega(w_0, 1) \models \neg occurs(\beta, 1) > occurs(escape, 1)$$

Giving negation narrow scope the left hand side is translated to:

$$occurs(unlock@obj(Rear); occurs(exit@obj(x), 1) \wedge x \neq Rear?, 1)$$

The counterfactual then follows directly. Finally, we must show that if the prisoner had known that exiting through the rear would not result in an escape, then the agent would not have exited in that way. This follows by minimizing unsupported beliefs: the conclusion from Definition 13.3 of unlocking and then exiting will no longer be triggered since the antecedent of that postulate is retracted by the contrapositive of Rationality Postulate 11.1. □

Theorem 15.3. "*Taking the rear exit was a mistake*".

Proof. This follows directly from the definition of a mistake since an alternative was possible: exiting through the front where no guard was present. □

Theorem 15.4. "*The presence of the guard prevented the escape*". *That is,*

$$occurs(guard(Rear)?, 1) \textbf{ prevented } occurs(escape, 1)$$

Proof. By Definition 6.1 for prevention, since we know that $occurs(guard(Rear)?, 1)$ and also that $holds(\neg\Diamond_\Omega Occurs(escape, 1), 1)$, we need only show that:

$$\sigma_\Omega(w_0, 1) \models \neg occurs(guard(Rear)?, 1) > holds(\Diamond_\Omega Occurs(escape, 1), 1)$$

We have that

$$\sigma_\Omega(w_0, 1) \diamond \|\neg occurs(guard(Rear)?, 1)\| = \sigma' =$$
$$\|\mathcal{WD}_\Omega(w_0, 1) - \{holds(guard(Rear), 1)\} \cup \mathcal{L}_\Omega(w_0, 1) \cup \mathcal{P}_\Omega(w_0, 1)\|$$

We then have that $min\{\sigma' \diamond \|\neg occurs(guard(Rear)?, 1), \preceq_E \|\} = \{w_1, w_2\}$ where in one of those worlds there is no guard and therefore an escape is possible. □

16. Implementation

16.1. Microworld description

This section describes a Prolog program that responds to queries involving causal relations between the activities of a collection of agents engaged in purposeful behavior within a micro-world. This implementation was meant to serve two purposes: to demonstrate the theoretical claims of this work and, more importantly, as a *problem generator* during the analysis process, that is, as a source of novel situations to motivate subsequent analyses and through which existing analyses might be refined. Such a research methodology—in which the theoretical analysis is grounded in concrete problem instances—seems preferable to one based simply on introspection. Approaches of the latter sort depend on one being sufficiently fortunate to have anticipated every possible circumstance the theory should explain.

The micro-world is embedded within an environment called the *Rational AGency Universe* (RAGU). RAGU is an environment for constructing simulations of groups of rational agents engaged in goal-directed behavior. In RAGU, one can specify the "physics" of the world by way of statements in a reified, horn-clause version of $\mathcal{HL}$. In this work, RAGU is used to implement a game called JASON which consists of a rectangular maze-like structure along which agents maneuver. JASON comprises three agents: Jason the argonaut who is in search of the golden fleece, a Dragon who continuously follows Jason around, and a Hydra which guards the golden fleece. Contact with either the Dragon or the Hydra results in Jason's destruction. To therefore assist Jason, hidden in the maze is a sword with which he can kill the hydra and a spear with which he can kill the dragon. In order to introduce uncertainty into this world, the perception of Jason is limited to a three by three grid around Jason. In contrast, the Dragon always knows where Jason is but may not be too smart in going after him.

16.2. Example scenario

Fig. 1 illustrates a sample scenario taken from JASON. The maze consists of stone walls which cannot be crossed and bodies of water which can only be crossed by dragons. Among other objects shown are the sword and spears already mentioned as well as a vase which contains a libation which will slow Jason down if he drinks from it. The shield can protect Jason from the Dragon (i.e., he can navigate directly past the Dragon without fear) and the oracle can be visited by Jason: it can supply him with either the current location of the dragon or the location of the fleece. The fleece is illustrated in this figure in the lower right hand corner. Also shown is a unicorn which Jason can mount in order to speed up (if it has consumed wine) or cross bodies of water.

In the scenario shown, Jason turns to the west along the path highlighted by the solid circles, picking up the sword along the way (recall that Jason can use the sword against the hydra but not against the dragon). When Jason arrives at the end of the path shown he takes the west fork until he gets to the bottom left-hand corner of the board. In the meantime, the dragon is following Jason and traps it at the bottom corner and the game ends. The board is 10×10 and the location of the various objects on the board is in terms of pairs, (X, Y),

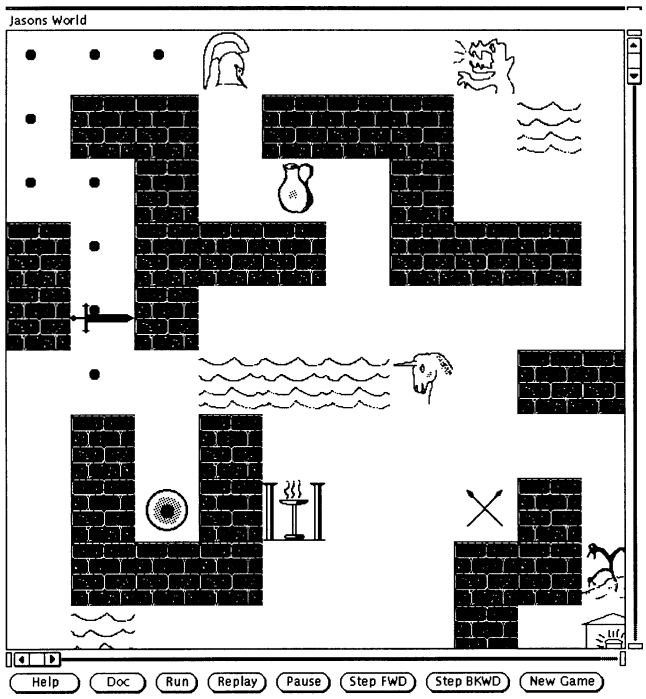

Fig. 1. Hypothetical scenario from RAGU microworld.

where X represents the number of the square along the X direction, and Y the number of the square along the Y direction, both starting at the top left-hand corner. For example, the initial location of Jason is $(4, 1)$ and the location of the fleece is $(10, 10)$.

The following series of queries and responses—posed at the point at which Jason arrives at point $(2, 6)$—is meant to elaborate the sorts of technical issues with which this work has been concerned. The actual queries are presented to the system in a formal language; natural language is being used here simply for expository purposes. All of the queries shown can be handled by the system.

```
U1> Did taking the west fork prevent Jason from escaping
    the dragon?
S>  Yes.
```

This exchange illustrates the semantics of prevention. In this case, if Jason had not taken the west fork it would have been possible to escape the dragon. After he took the west fork, however, that was no longer possible.

```
U2> Did moving away from the dragon cause Jason to be
    destroyed?
S>  No.
```

This is an example of the non-truth functional nature of the causal relation (what has been called the event subsumption problem in this paper). It is true that taking the west

fork caused Jason to be destroyed—because if he had taken the east fork, he would have been saved by picking up the shield.[25] However, in the system there is, as one would expect, no explicit rule with such a description as an antecedent. Instead there is only knowledge present involving the effects of coming into contact with the Dragon, together with other non-causal knowledge such as a definition for what it means for an agent to move away form some object, what it means for one event to occur before another, and so forth.

```
U3> Would Jason have been destroyed if he had taken
    the east fork?
S>  No.
```

It follows that, by exploring the counterfactual worlds associated with Jason taking the east fork instead, we cannot conclude that he would have been destroyed in *all* of those worlds: there is a world in which he picks up the shield and is protected.

```
U4> Could taking the east fork have enabled Jason to
    escape the dragon?
S>  Yes.
```

The possibility that Jason could have escaped if he had taken the east fork is equivalent to saying that taking the east fork would have enabled him to escape. As discussed earlier, the notion of enablement is weaker than that of causation.

```
U5> Did picking up the sword before moving on cause
    Jason to be destroyed?
S>  No.
U6> (At the point at which Jason took the west fork)
    Did not moving towards the dragon cause Jason's
    destruction?
S>  No.
```

Both of these queries are also examples of the event subsumption problem. Alternative descriptions of an event are often formed by way of temporal or spatial adverbs: for example, picking up the sword *before* moving on, picking up the sword instead of picking up something else, etc. Again, even though the description of the event contained in U5 is a true description, it is inadequate and in fact false as an explanation of the caused event in question. In U6, a *negative event description* is made use of. The falsity of the report once again falls out of a counterfactual analysis: if he had moved towards the dragon he would not have been saved.

The following two queries are made at the point at which the dragon is within Jason's range of perception (i.e., at the bottom, left-hand corner of the board).

```
U7> Did Jason try to escape when he saw the Dragon?
S>  Yes.
```

[25] Jason will automatically pick up the shield if he can.

```
U8> Did he succeed?
S>  No.
```

These exchanges are representative of reports of *attempts* and *failures*. An attempt can fail for various reasons: for example, because of a mis-match between an agent's beliefs and the true facts of a situation. Notice that if this query had been posed at the point at which Jason took the west fork, it would have been false since Jason had no idea that the dragon was following him at that point and so did not intend to get away. This intentional precondition is necessary for an escape.

```
U9>  Was taking the west fork a mistake?
S>   Yes.
U10> Was taking the west fork an accident?
S>   No.
```

Related to failures are accidents and mistakes. Mistakes represent incorrect choices in trying to perform some ends action by some means action. In the example, the agent intended to escape by taking the west fork, but he was incorrect in his choice of means action—taking the west fork—which did not represent a *method* of performing the indicated higher-level action.

```
U11> If the Unicorn had been at (1,10), would riding the
     unicorn have helped Jason  escape the dragon?
S>   Yes.
```

In this case, if the Unicorn had been located at the bottom left-hand corner of the board, Jason could have used it to cross the body of water adjacent to that point and escaped the dragon.

```
U13> Did the dragon let Jason take the west fork?
S>   No.
```

At the moment when Jason took the west fork, it was not possible for the dragon to prevent him from doing so.

```
U14> Did the dragon kill Jason by touching him?
S>   Yes.
U15> Did the dragon's touch cause the dragon
     to kill Jason?
S>   No.
```

A direct method-of relation exists here between touching Jason and killing Jason. Again, since a counterfactual relation exists between the touching and the killing, some additional conditions are imposed to distinguish such cases from bona-fide cases of causation: one must block cases such as `U15`.

```
U16> If Jason had taken the first path could the dragon
     have kept him away from the fleece?
S>   No.
```

That is, if Jason had moved south immediately from its starting position—toward the vase—could the dragon have kept him away from the fleece? This is an example of a *maintenance action*. It is essentially asking whether there was some series of reactions on the part of the dragon to actions on the part of Jason where the latter were directed towards getting to the fleece.

16.3. Representation and reasoning

The event representation is as faithful as possible to a reified version of $\mathcal{HL}$, with the exception of a restriction to horn clauses and negation by way of the closed world assumption. To report that some fact p held in world w at time t, one writes the prolog statement `holds(p,w,t)` and to state that some event, e occurred at time t in world w, one writes `occurs(e,w,t)`. Dynamic logic connectives are introduced as in $\mathcal{HL}$. The ones made most use of are sequencing, and tests:

```
occurs((E1;E2),W,T1) :-
        occurs(E,W,T1),
        sub_type(E1,E),
        extract(dur,E,D),
        T2 is T1 + D,
        occurs(E2,W,T2).
occurs(P?,W,T) :-
        holds(P,W,T).
```

The physics of the game is specified in terms of causal rules giving preconditions and effects for actions, as well as additional definitional knowledge. For example, the pickup action is defined as:

```
holds(has(A,X),W,T) :-
      dur(T,T1,1),
      fact(occurs(pickup@agt(A)@obj(X),W,T1)),
      holds(cond(grabbing@agt(A)@obj(X),has(A,X)),
            W,T1).
```

where a separate definition is given for the last clause specifying the preconditions for the action to have the indicated effects: that the agent be at the same location as the object.

Control is specified by a think-act loop which takes care of:

(1) marking locations as visited,
(2) calling on a perception routine for Jason and the dragon which asserts which objects each is aware of,
(3) performing the indicated actions and recording the effects of the perceptual actions,
(4) incrementing the state,
(5) forming intentions to perform actions by considering possible actions at the current state that could achieve the agent's goals,
(6) choosing an action to perform through a very simple deliberation routine, and
(7) performing the physical actions and recording the effects.

Intentions are represented by the system in reified form also: Jason's intention at time t to perform some α is recorded as $holds(Int(Jason, \alpha), t)$. For example, initially Jason has the intention of getting to the fleece; if he knows the direction of the fleece he will also form the intention to move in that direction. In another situation, if he sees the dragon, he will also form the intention to get away.

Two routines for evaluating counterfactuals were developed: one was based directly on the information change operations described earlier. The second made some simplifying assumptions in order to focus on the testing of the various causal relations: counterfactual worlds were explored by a simple traversal of a branching tree. In the second approach, however, preferences on possible worlds *were* encoded in order to compute, for example, the consequences of some event *not* occurring. Here, only the first, more general scheme will be described. In that scheme, counterfactuals are defined in terms of the following revision procedure

```
revise(+Order,+World,+Time,+Fact,-NewWorlds),
```

where a theory defined at `World` is revised with `Fact` relative to some ordering, `O`. There are three cases. The first two cases are the easiest:

(1) `Fact` is already part of the current theory at `World` and `Time` and therefore nothing needs to be done in this case, or

(2) `Fact` does not conflict with the current theory in which case it can simply be added.

The last case is the interesting one where `Fact` conflicts. In that case, the minimal revision is made. In each case we need to create new worlds:

```
revise(O,W,T,P,Ws) :-
       negate(P,Q),
       retract(Q,W,T),
       min_revision(O,W,T,P,Ws).
```

The predicate

```
min_revision(+O,+W,+T,+P,-NewWorlds)
```

makes the minimal revision according to the partial order. It does so by first collecting all possible proofs of `P` and then for each such proof, picking the smallest element(s) according to the ordering. `MinElmnts` is a list of lists—one list for each possible proof. It then takes the product of these minimal elements and creates new worlds for each new knowledge base with those elements deleted.

```
min_revision(Order,W,T,P,Ws) :-
       negate(P,Q),
       clause(holds(Q,W,T),true),
       setof(C,Q^W^T^prove(holds(Q,W,T),C),Proofs),
       filter_facts(Proofs,Ps),
       max_elements(Order,Ps,MaxElmnts),
       product(MaxElmnts,[],DelElmnts),
       create_new_worlds(W,T,P,DelElmnts,Ws).
```

The predicate `prove/2` calls on a meta-interpreter which maintains justifications and associates negation with "clipped" fluents, written `clipped(Fact,Event)` in prolog, and meaning that `Fact` has been negated by `Event`. Here is a sample clause (the second argument stores the assumptions at the leafs of the proof for the given facts):

```
prove(holds(P&Q,W,T),A) :-
       prove(holds(P,W,T),A1),
       prove(holds(Q,W,T),A2),
       append(A1,A2,A).
```

Some examples of definitions of act-types include:

```
occurs(try(Act),W,T) :-
       ext(agt,Act,Agent),
       holds(intends(Agent,by(Means,Act)),W,T),
       occurs(Means@agt(Agent),W,T).

occurs(not(Act)@agt(Agent)@obj(_)@manner(fails),W,T):-
       occurs(try(Act@agt(Agent)),W,T),
       \+occurs(Act,W,T).
```

As an example of the form that the actual queries take, here is a sample query and response in the representation language of RAGU:

```
% Did moving away from the dragon cause Jason
  to be destroyed?
% This query demonstrates the modal nature
  of the causal connective.
% One cannot freely substitute descriptions
  of other true events.

| ?- holds(causes(move@agt(jason)@away(dragon),
             future(clipped(alive(jason),
             touch@agt(dragon)@obj(jason)))),
             w0,50).
no
```

17. Summary

A commonsense language for reasoning about causation and rational action that is counterfactually based has been presented. Beginning with a semantics for counterfactuals based on belief updating, a semantics of causation was developed; the resulting formulation ruled out cases in which method-of or part-of relations between actions might be incorrectly identified as genuine causal connections. In addition, the counterfactual dependency which underlies the definition of causation was given in terms of antecedent

Table 1
Summary of the general characteristics of the relations examined. The pragmatic characteristics in the third column correspond to questions of appropriateness of the various relations in a causal explanation and were only discussed in very general terms

Causal relation	General characteristics	Pragmatics of explanation
Enables β	$\Diamond\beta$ after α	α background if $\Box\beta$
Causes β	$\Box\beta$ after α; α instrumental	α foreground; β accident
Forces β	More resources to cause β	
Prevents β	β never occurs after α	
Maintains ϕ	a process; prevents $\neg\phi$	
Helps β	α reduces resources for part_of β	
Hinders β	α helps increase resources	
Lets β	β not prevented and not caused	
α method β	β higher-level act	β "by" α

classes of event types. This was important in order to solve the problem of pre-emption or causal overdetermination.

Further refinements of the theory that could accommodate the observed differences among the counterfactual dependencies underlying commonsense causal language were then considered. Examples included notions of prevention, enablement, causation, and coercion. The utility of reasoning about counterfactual worlds to the task of causal explanation in which the proper choice of event description was critical to a proper analysis was demonstrated; in particular to causal reports involving negative event descriptions.

Finally, an implementation of a fragment of the above formalism was described: the domain consisted of a microworld of agents involved in goal-directed behavior. The goal of this part of the research was to test and evaluate the ideas and definitions developed as part of the commonsense language and not as a means of implementing the entire theory. In particular, the utility of counterfactual reasoning in dealing with the event subsumption problem was demonstrated by considering the acceptability of a variety of causal reports between alternative action descriptions of the same event.

Table 1 summarizes the major causal connections examined and Table 2 summarizes the various act-types examined.

17.1. Limitations and future work

What can be said as a step in generalizing the properties essential to differentiating the various causal relations? Assuming that $REL(X, Y)$ represents some causal relation between act-types X and Y, then each relation can be defined in terms of a set of possible features, where each feature is given a value for the real world and one for the counterfactual world. These features include: cost (number of resources, physical force, etc.); aspectual classification (an event, process, etc. [6]); spatio-temporal profile (as in Rieger [60]: whether the causing event continues to "impinge" on the caused event during

Table 2
Summary of the act types investigated. Only the general characteristics of each act-type are summarized

Act-type	General description
Agentive	Grounded in basic act
Intentional	Intended or believed side-effect
Accident	Ends act: wouldn't have done it otherwise
Mistake	Means act: wrong choice in trying to ends
Attempt	Intended to β by some α and α-ed
Failure	An attempt where wrong α or wrong beliefs

the range of interaction or the physical participants in the causal relation and their states at various times); and whether or not the causing event was instrumental. This is left for future work.

Other areas that have not been addressed in this paper are the following:

- Explanation at the pragmatic level: that is, identifying the salient cause among a number of alternative causal factors.
- A taxonomy of negative event types such as cases of avoiding, omitting, and refraining.
- Causal relations between states of the world (and not events). For example, although a case such as "the presence of the flammable gas caused the fire" would be handled correctly, statements such as "the legs on the table caused it to have a height of three feet" are not handled properly.
- The role of aspectual composition in causal explanations (whether some event is a process, point event, or non-homogeneous).
- The acquisition of the sorts of causal knowledge as related to the sorts of primitive relations discussed in this paper.

Acknowledgement

This paper, together with a companion paper [58], represents revised dissertation work [57]. The interested reader is referred to the companion paper for details on EUT as well as counterfactual reasoning. Special thanks to Mark Steedman, Leora Morgenstern, and anonymous referees for comments on earlier versions of this paper. This work was supported by a University of Pennsylvania Fontaine Fellowship, grants from ARO, and NSF grant No. IRI 95-25915.

Appendix A. Syntax and semantics of $\mathcal{HL}$

This section presents the syntax and semantics of *hypothetical logic* ($\mathcal{HL}$). The language is a sorted modal logic with sorts for event types, objects, times, and fluents.

A.1. Syntax of $\mathcal{HL}$

Definition A.1. The ALPHABET for the language consists of:

(1) a set of constants $\mathcal{C} = C_t \cup C_d \cup C_\tau$ where C_t is a set of event-type constants, C_d is a set of object constants, and C_τ is a set of time constants (integers). C_t includes ε, for the null event;
(2) a set, $V = V_t \cup V_d \cup V_\tau \cup V_{Fluents}$, of object variables, one for each sort;
(3) two binary predicates: *holds* ranging over propositional fluents and time and *occurs* ranging over action-terms and time;
(4) a set, $\mathcal{F} = F_t \cup F_d \cup Fluents$, of functions on C_t, C_d, and C_d, respectively, together with an event-type constructor $@ \in F_t$;
(5) quantification, $\forall$;
(6) logical connectives $\{\neg, \supset\}$;
(7) two modalities, $\Box_W$, $\Diamond$, and a counterfactual connective, $>$.

The truth of a formula will be given relative to some world. In $\mathcal{HL}$, time is reified and discrete; that is, time ranges over the integers. To report that ϕ held at time point t (relative to the current world), we write: $holds(\phi, t)$ and to report that act-type α occurred at time t, we write $occurs(\alpha, t)$, again relative to the current world. The duration of an event will be specified in the description of that event by way of an event-type constructor which creates complex event-types from simpler ones: for example, $pickup@agt(John)@manner(slowly)@dur(10)$ might name the event type of some agent, John, picking up an object slowly, that event having taken 10 time units. The connective $>$ is the usual one for counterfactual dependence: $\phi > \psi$ is meant to express the fact that if ϕ had been (instead) true, then so would have ψ. Finally, $\Box_W \phi$ will mean that ϕ is true in all possible worlds (at the given time), [26] and $\Diamond\alpha$ means that there is some physically possible world in which α is performed.

Definition A.2. The set of NON-ACT TERMS is the smallest set, $\mathcal{T}$:

(1) if $t \in C_d \cup C_\tau \cup V_d \cup V_\tau$ then $t \in \mathcal{T}$; and
(2) if $t_1, \ldots, t_n \in \mathcal{T}$ and $f \in F_d \cup Fluents$, with arity n, then $f(t_1, \ldots, t_n) \in \mathcal{T}$.

Definition A.3. The set of ACT-TYPE TERMS is the smallest set, $\mathcal{A}$, such that:

(1) if $e \in C_t \cup V_t$ then $e \in \mathcal{A}$; and
(2) if $e \in \mathcal{A}$ and $f \in F_d, x \in TERMS$ then $e@f(x) \in \mathcal{A}$.

An event type will be taken to be coextensive with the set of intervals in which it occurs.

Definition A.4. The set of terms, *TERMS*, is $\mathcal{T} \cup \mathcal{A}$.

Definition A.5. The set of formulas is the smallest set, *FORMS*, such that:

(1) if $x \in Fluents$ and $t \in C_\tau$ then $holds(x, t) \in FORMS$;
(2) if $\alpha \in \mathcal{A}$ and $t \in C_\tau$ then $occurs(\alpha, t) \in FORMS$;

[26] The W subscript is not an index to the $\Box$ operator; the pair represents a single symbol.

(3) if $p, q \in FORMS$ then so are $p \supset q$ and $\neg p$;
(4) if $v \in V$ and $\phi \in FORMS$, then $\forall v.\phi \in FORMS$;
(5) if $\phi \in FORMS$ then $\Box_W \phi \in FORMS$;
(6) if $\phi \in FORMS$ then $\Diamond\phi \in FORMS$;
(7) if $\phi, \psi \in FORMS$ then so is $\phi > \psi$.

The usual additional logical connectives ($\wedge, \vee, \equiv$) are introduced by definition, as is the existential quantifier: $\exists x.\phi =_{def} \neg\forall x.\neg\phi$.

A.2. Semantics of $\mathcal{HL}$

The semantics for $\mathcal{HL}$ is given in terms of *Kripke Structures* [15,32]. A *model*, $\mathcal{M}$, is defined as a structure, $\langle D, W, T, \sqsubseteq_w, A_i, I, v\rangle$. $D = \{D_d, D_e, Agt\}$ represents a domain of individuals common to each possible world, consisting of physical objects, D_d, and agents, *Agt*. Formulas in the language are interpreted relative to some model, world, and time, where W is a set of possible worlds, and T a set of times, here taken from the set of integers; that is, models are restricted to discrete time. Associated with each possible world, agent, and time is a set of formulas or *assumptions*, A, such that $A : Agt \times W \times T \rightarrow FORMS$. Worlds, w, are ordered according to a partial ordering, $\sqsubseteq_w$; one for each world, w. This ordering reflects relative similarities between possible worlds in a way determined by the set of assumptions, A. Roughly speaking, world u is $\sqsubset_w$-closer to world w, relative to some set of assumptions, $A(w,t)$, than world v is (written $u \sqsubset_w v$) if u has more of the initial set of assumptions, $A(w,t)$, than v does. Formulas and terms are interpreted by way of an interpretation function, I, and a variable valuation, $v : V \rightarrow D$. The semantics for the $\Diamond$ operator will be given in section 17.1. The semantics for the counterfactual connective will also be discussed in the next section.

In this language, the interpretation of an event type is the set of world-time interval pairs over which it occurs. It is therefore useful to be able to refer to arbitrary spans of time within a world.

Definition A.6. Let the set of *world-time intervals*, $\overline{T}$, be defined as:

$$\overline{T} = \big\{\langle w, t_j, \ldots, t_n\rangle \mid w \in W, t_i \in T, \text{ for } j \leqslant i \leqslant n, \text{ such that, } t_{i+1} = t_i + 1\big\}.$$

If an event has duration of one time unit and occurs in world w and time t, then this means that it will occur over the interval $\langle w, t, t+1\rangle$. If an event occurs at time t in world w and has no duration then it occurs over the interval $\langle w, t\rangle$. Examples of events with zero duration include the distinguished null event as well as test events. Point events representing occurrences such as "the ball reached its highest point (after being thrown in the air)" can be modeled instead as actions testing the truth of "highest point". The representation is neutral, however, with respect to many important ontological questions [35].

Definition A.7. Let $I = (I_d, I_t, I_\tau, I_f, I_p)$. Then the interpretation of terms is given by:
(1) $I_d : C_d \rightarrow D_d$;
(2) $I_t : \mathcal{A} \rightarrow \overline{T}$;

(3) $I_\tau : C_\tau \to T$;
(4) $I_f : (\mathcal{F} - \mathit{Fluents}) \to (W \times T \to (D^n \to D))$; and,
(5) $I_p : \mathit{Fluents} \times D^n \to 2^{W \times T}$.

Definition A.8. The denotation of a term, s, relative to some model, $\mathcal{M}$, world, w, and time, t, written as $[\![s]\!]_t^w$ (where reference to $\mathcal{M}$ is suppressed when understood):
(1) if $c \in C_x$ where $x \in \{d, t, \tau\}$ then $[\![c]\!]_t^w = I_x(c)$;
(2) if $u \in V$ then $[\![u]\!]_t^w = v(u)$;
(3) $[\![f(r)]\!]_t^w = I_f(f)(w, t)[\![r]\!]_t^w$.

In this semantics, constants are rigid designators whereas functional expressions are not. As Shoham and others have shown, this is useful when constructing terms that depend on a temporal component (e.g., *president*(*usa*)) [62].

Let $\phi[x/d]$ stand for the substitution instance of d for x in ϕ.

Definition A.9. A formula ϕ is *satisfiable* in M at w, written $M, w \models \phi$, just in case it is subsumed by one of the following cases:
(1) if $p \in \mathit{Fluents}$, of arity n, then $M, w \models \mathit{holds}(p(t_1, \ldots, t_n), t)$ iff $\langle w, t\rangle \in I_p(p)$ $([\![t_1]\!]_t^w, \ldots, [\![t_n]\!]_t^w)$;
(2) if $\alpha \in \mathcal{A}$ then $M, w \models \mathit{occurs}(\alpha, t)$ iff $\exists \gamma = \langle w, t, t+1, \ldots, t+n\rangle$ such that, $\gamma \in [\![\alpha]\!]_t^w$;
(3) if $\phi, \psi \in \mathit{FORMS}$ then $M, w \models \phi \supset \psi$ iff either $M, w \models \psi$ or $M, w \not\models \phi$;
(4) if $\phi, \psi \in \mathit{FORMS}$ then $M, w \models \neg\phi$ iff $M, w \not\models \phi$;
(5) $M, w \models \forall x.\Phi(\phi, t)$ iff for all $d \in D$: $\langle w, t\rangle \in I_p(\phi[x/d])$, where $\Phi \in \{\mathit{holds}, \mathit{occurs}\}$;
(6) $M, w \models \Box_W \phi$ iff for all w': $M, w' \models \phi$.

In addition, the common meta-theoretic definitions: $\mathit{holds}(\phi, t) \wedge \mathit{holds}(\psi, t)$ just in case $\mathit{holds}(\phi \wedge \psi, t)$, etc will be used.[27] If ϕ is satisfiable in all models (i.e., valid), we write $\models \phi$. Here is an example of a formula in the language:

$$w \models \mathit{holds}(\mathit{on}(A, B), t) \wedge \mathit{occurs}(\mathit{put@agt}(\mathit{Harry})@\mathit{obj}(C)@\mathit{on}(A)@\mathit{dur}(5), t)$$

This reports that in world w, time t, block A was on block B and Harry carried out the action having duration 5 of putting block C on A.

In $\mathcal{HL}$, possible worlds are viewed as alternative possibilities from the point of view of the language used to describe worlds—that is, maximally consistent sets of formulas [43]. This is in contrast to the realist position of Lewis [45]. In the above, the domain, D, is common to all possible worlds. This raises the usual question of how to deal with statements involving individuals that might not exist in some worlds [27]. In the context of counterfactual reasoning, this arises in examples such as the following:

(46) If we had not won the revolutionary war, then George Washington would not have been the first president.

[27] Actually, the logical operators that occur inside of the holds statements are really functions. However, the same symbol will be used for both.

In this example, the *first president* is an individual that simply might not have existed if the revolutionary war had not been won. This sort of example can be handled through the inclusion of an *existence predicate*, E (actually, a term in this reified language), where $I_f(E) : D \to 2^{W \times T}$. In this way, every formula receives a truth value. The following axiom relates the existence predicate to domain closure axioms:

Rationality Postulate A.1 (*Existence in alternative worlds*).

$$\models \forall x.\mathit{holds}(E(x), t) \equiv \exists y.\mathit{holds}(y = x, t) \tag{A.1}$$

In $\mathcal{HL}$, basic actions will be assumed to always succeed. This represents no real problem as one can allow the set of basic actions associated with an agent to vary with time [57].

Fig. A.1 axiomatizes the @ event-type constructor using relativized quantifiers. The @-constructor essentially constructs a set consisting of an event type, α, and a set of modifiers, $x_1, x_2, \ldots, x_n$, separated by @. The language used consists of Λ, the empty collection of modifiers; $\mathit{atom}(x)$, denoting the *atom* relation true if x is a single element; $\mathit{mods}(x)$, a predicate true if x is nonatomic, that is, is some set of modifiers $\alpha@x_1@x_2@\cdots@x_n$; $\mathit{in}(x, y)$, denoting the membership relation; and a predicate $\mathit{sub\_type}(x, y)$. A set of names of event types is distinguished by way of the *name* predicate; more complex descriptions of event types are formed by appending modifiers using the @ function. Terms such as $\alpha@\Lambda$ are written simply as α. Axiom 6 from the figure allows one to conclude that the following are equivalent:

$$\mathit{pickup@agt}(\mathit{John})@\mathit{obj}(\mathit{Block})@\mathit{agt}(\mathit{John}) = \mathit{pickup@agt}(\mathit{John})@\mathit{obj}(\mathit{Block})$$

Axiom 7 says that the following two event descriptions are equivalent names for the event of agent John picking up a block with his right hand:

$$\mathit{pickup@agt}(\mathit{John})@\mathit{obj}(\mathit{Block})@\mathit{with}(\mathit{Right\_hand}) = \mathit{pickup@obj}(\mathit{Block})@\mathit{with}(\mathit{Right\_hand})@\mathit{agt}(\mathit{John})$$

1. $\models \mathit{holds}(\mathit{event\_type}(\alpha@m) \equiv \mathit{name}(\alpha) \wedge \mathit{mods}(m), t)$
2. $\models \mathit{holds}(\mathit{mods}(\Lambda), t)$, the empty modifier
3. $\models \forall \mathit{atom}(u)\forall \mathit{mods}(m).\mathit{holds}(\mathit{mods}(u@m), t)$, groups of modifiers
4. $\models \forall \mathit{atom}(u)\neg \mathit{holds}(\mathit{in}(u, \Lambda), t)$
5. $\models \forall \mathit{atom}(u)\forall \mathit{atom}(v)\forall \mathit{mods}(m).\mathit{holds}(\mathit{in}(u, (v@x)) \equiv u = v \vee \mathit{in}(u, x), t)$, membership
6. $\models \forall \mathit{atom}(u)\forall \mathit{mods}(x).\mathit{holds}(u@(u@x) = u@x, t)$, irrelevance of copies
7. $\models \forall \mathit{atom}(u)\forall \mathit{atom}(v)\forall \mathit{mods}(x).\mathit{holds}(u@(v@x) = v@(u@x), t)$, order irrelevance
8. $\models \forall \mathit{mods}(y).\mathit{holds}(\mathit{sub\_type}(\Lambda, y), t)$, sub-modifiers
9. $\models \forall \mathit{atom}(u)\forall \mathit{mods}(x)\forall \mathit{mods}(y).\mathit{holds}(\mathit{sub\_type}(u@x, y) \equiv (\mathit{in}(u, y) \wedge \mathit{sub\_type}(x, y)), t)$

Fig. A.1. Axiomatization of the @ event-type constructor.

Appendix B. Information change and the semantics of counterfactuals

The semantics for counterfactuals adopted here is one based on belief update [39,75] called Explanatory Update Theory (EUT) [57,58]. Consider a mapping of each $u \in W$ to a partial pre-order, $\sqsubseteq_u$, such that the following condition is satisfied: for any $v \in W$, if $u \neq v$ then $u \sqsubset_u v$.[28] Given some set of worlds, V, the notation $min(V, \sqsubseteq_u)$ stands for the set of $w \in V$ such that w is minimal in V with respect to $\sqsubseteq_u$. Of course, the interesting cases are those in which $u \notin V$. Katsuno and Mendelzon place no other constraints on $\sqsubseteq_u$ (but see below). The *update* of some theory, T, at point $\langle w, t\rangle$, with some ψ is then defined as [39]:

$$\|T(w,t)\| \diamond \|\psi\| =_{def} \bigcup_{u \in \|T(w,t)\|} min\big(\|\psi\|, \sqsubseteq_u\big)$$

That is, to update some theory, T, with some ψ, we take the union of all the $\sqsubseteq_u$-minimum ψ-worlds for each T-world.

In order to deal with the frame problem, EUT specifies how an agent's *information state*, σ, can be computed: σ stands for the set of possible worlds that include all of the agent's initial assumptions (that is, T) together with whatever additional assumptions of persistence are necessary in order to deal with the frame problem. This is accomplished through an extension of the ideas in Motivated Action Theory (MAT) of Stein and Morgentstern [65]: worlds are minimized according to an ordering, $\preceq_E$, which assigns lower preference to beliefs that have no support from other beliefs; that is, beliefs that cannot be "explained". For the purposes of this paper the details of the construction are not important: it is sufficient to assume some mechanism through which this can be accomplished. Counterfactual dependence can then be defined as follows. First, the notion of satisfaction is extended to sets of worlds: $U \models \phi$ just in case $u \models \phi$ for each $u \in U$. When some formula is evaluated relative to a single world—as in $w \models \phi$, this will be taken to mean $\{w\} \models \phi$, with the same meaning as before. We then have,

$$\sigma(w,t) \models occurs(\alpha, t_1) > occurs(\beta, t_2) \text{ iff}$$
$$min\{\sigma(w,t) \diamond \|occurs(\alpha, t_1)\|, \preceq_E\} \models occurs(\beta, t_2)$$

The above says that $\phi > \psi$ (read: "if ϕ had been the case, ψ would have been the case") is true relative to some σ at $\langle w, t\rangle$, just in case ψ holds in all worlds in which the ϕ-updated information state has been explained. EUT then describes the sorts of extra-logical constraints that should be place on $\sqsubseteq_u$ (relative to the set of assumptions given for world u) in order to support this definition and application domain:

(1) causal knowledge should be protected;
(2) worlds in which the objects referred to by the counterfactual supposition exist are preferred; and
(3) preference is given to localized differences in accommodating a counterfactual supposition (else, one might unravel the past up to some branch point that represented a common cause of some other, unrelated fact).

The interested reader is referred to the companion paper for details on EUT.

[28] Recall that a preorder is a reflexive and transitive relation.

Belief reports are given a sentential semantics in terms of the agent's information state. First, since belief statements take formulas as arguments, we need a functional term to stand for holds statements: the notation $Holds(\phi, t)$ [29] will be used. An agent's belief at time t in some ϕ, itself true at time t', will be written as $holds(Bel(i, Holds(\phi, t')), t)$. The semantics of belief reports is then given as follows.

Definition B.1 (*Belief in terms of information state*).

$$w \models holds(Bel(i, Holds(\phi, t')), t)$$
$$\text{iff } \sigma_i(w, t) \models holds(\phi, t')$$
$$\text{iff } min\{\|\mathcal{A}_i(w, t)\|, \preceq_E\} \models holds(\phi, t')$$

$\mathcal{A}_i(w, t)$ represents i's set of assumptions at $\langle w, t\rangle$; this set will generally be equated with the union of a set of laws or nomic expressions, $\mathcal{L}_i(w, t)$ and an initial world description, $\mathcal{WD}_i(w, t)$. The above says that at time t agent i believes that ϕ holds at time t' just in case $holds(\phi, t')$ is true in all worlds given by the agent's information state. Other statements embedded within a belief term will be handled in a similar way. For instance:

$$w \models holds(Bel(i, Occurs(\alpha, t')), t) \text{ iff } \sigma_i(w, t) \models occurs(\alpha, t')$$

Knowledge can be distinguished in the usual way:

$$\models holds(Know(i, Holds(\phi, t')), t) \equiv holds(Bel(i, Holds(\phi, t')), t) \wedge holds(\phi, t')$$

Konolige investigated sentential approaches thoroughly including the conditions under which they correspond to well known possible worlds epistemic logics [42].

An epistemic notion of possibility [57] can be defined as follows:

Definition B.2 (*Possibility in terms of information state*).

$$w \models holds(\Diamond_i Occurs(\alpha, t'), t) \text{ iff}$$
$$\exists w' \in \sigma_i(w, t) \text{ such that, } w' \models occurs(\alpha, t')$$

and similarly for the case of propositional fluents:

$$w \models holds(\Diamond_i Holds(\phi, t'), t) \text{ iff}$$
$$\exists w' \in \sigma_i(w, t) \text{ such that, } w' \models holds(\phi, t')$$

In cases of formulas of the form $holds(\Diamond_i Holds(\phi, t'), t)$, any propositional fluent with embedded logical connectives under the scope of the $\Diamond_i$ modality will appear in the ϕ term; that is, in statements of the form: $holds(\Diamond_i Holds(\phi \wedge \psi, t'), t)$, and so forth. Cases of embedded $\Diamond$'s or *Bel* operators are not considered in this paper. To distinguish actual possibility from epistemic possibility, it is often useful to posit some omniscient agent, here referred to as Ω, whose beliefs are correct though not necessarily complete.

[29] Notice the use of upper case.

1. $\models occurs(\varepsilon@dur(0), t)$, the null event
2. $\models holds(basic(i, \alpha) \supset \Diamond(Occurs(\alpha@agt(i), t)), t)$, success of basics
3. $\models holds(part\_of(\alpha, \gamma; \alpha; \delta) \wedge part\_of(\varepsilon, \alpha), t)$, part-of
4. $\models occurs(achieve@obj(\phi), t) \equiv \exists\alpha.occurs(\neg\phi?; \alpha; \phi?, t)$, achievement events
5. $\models holds(f(\alpha, x), t) \equiv \exists y \exists z.\alpha = y@f(x)@z$, extraction
6. $\models occurs(e, t) \wedge holds(sub\_type(\alpha, e), t) \supset occurs(\alpha, t)$, modifier dropping
7. $\models \forall i.holds(basic(i, \alpha) \supset name(\alpha), t)$, basics
8. $occurs(\phi?, t) \equiv_{def} holds(\phi, t)$, test action
9. $occurs(\alpha \cap \beta, t) \equiv_{def} occurs(\alpha, t) \wedge occurs(\beta, t)$, parallelism
10. $occurs(\alpha \cup \beta, t) \equiv_{def} occurs(\alpha, t) \vee occurs(\beta, t)$, non-determinism
11. $occurs(\alpha; \beta, t) \equiv_{def} holds(dur(\alpha, d), t) \wedge occurs(\alpha, t) \wedge occurs(\beta, t + d)$, sequencing
12. $occurs(\alpha^*, t) \equiv_{def} occurs(\varepsilon, t) \vee (holds(dur(\alpha, d), t) \wedge occurs(\alpha, t) \wedge occurs(\alpha^*, t + d))$, iteration
13. $occurs(\text{WHILE } \phi \text{ DO } \alpha, t) \equiv_{def} occurs((\phi?; \alpha)^*; \neg\phi?, t)$
14. $occurs(\text{IF } \phi \text{ THEN } \alpha \text{ ELSE } \beta, t) \equiv_{def} occurs((\phi?; \alpha) \cup (\neg\phi?; \beta), t)$

Fig. B.1. Supplementary axioms and definitions.

1. $\models occurs(\alpha@dur(d); \beta, t) \equiv occurs(\alpha@dur(d); occurs(\beta, t + d)?, t)$
2. $\models occurs(\phi?; \psi?, t) \equiv holds(\phi \wedge \psi, t)$
3. $\models occurs(\alpha@dur(d), t) \equiv occurs(\alpha; occurs(\alpha, t - d)?, t)$
4. $\models \alpha = \beta \equiv \Box_W[\forall t.occurs(\alpha, t) \equiv occurs(\beta, t)]$

Fig. B.2. Some useful theorems in $\mathcal{HL}$.

Fig. B.1 lists some useful axioms and definitions. Axiom 5 from the figure extracts some property from an action description. For example, to extract the duration of an event,

$$holds(dur(\alpha@agt(i)@dur(4)@manner(slow)@obj(x), 4), t)$$

Axiom 6 allows one to conclude the former from the latter of the following pair:

$$occurs(pickup@agt(John)@obj(Block), T)$$

$$occurs(pickup@agt(John)@obj(Block)@with(Right\_hand), T)$$

Nomic expressions will be written in the following form.

Assumption B.1 (*Form of nomic expressions*). Causal laws in $\mathcal{L}_i(w, t)$ take the form:

$$occurs(\beta, t) \wedge occurs(\alpha, t) \supset holds(\psi, t')$$

where $t' \geqslant t$ and where β can be a test action, $\phi?$

Such expressions are assumed to contain only a single consequent; conjunctive effects are expressed as two rules. This is necessary because the syntactic form of causal rules plays a role in the evaluation of counterfactuals [58]. [30]

[30] The cited reference makes use of defeasible nomic expressions through the introduction of abnormality predicates in the form given above. These are not made use of in this paper.

References

[1] J.F. Allen, Towards a general theory of action and time, Artificial Intelligence 23 (1984) 123–154.
[2] G.E.M. Anscombe, Intention, Cornell University Press, Ithaca, NY, 1963.
[3] J.L. Austin, Ifs and cans, in: Philosophical Papers, Oxford University Press, Oxford, 1961.
[4] J.L. Austin, A plea for excuses, in: Philosophical Papers, Oxford University Press, Oxford, 1961.
[5] J.L. Austin, Three ways of spilling ink, in: Philosophical Papers, Oxford University Press, Oxford, 1961.
[6] E. Bach, The algebra of events, Linguistics and Philosophy 9 (1986) 5–16.
[7] C.T. Balkanski, Modeling act-type relations in collaborative activity, Technical Report, Harvard University, Cambridge, MA, 1990.
[8] C.T. Balkanski, Actions, beliefs, and intentions in multi-action utterances, Ph.D. Dissertation, Harvard University, Cambridge, MA, 1993.
[9] J. Baron, I. Ritov, Protected values and omission bias, Organizational Behavior and Human Decision Processes 59 (1998) 475–498.
[10] J. Bennett, Killing and letting die, in: S.M. McMurrin (Ed.), The Tanner Lectures on Human Values, Vol II, University of Utah Press, 1981, pp. 47–72.
[11] J. Bennett, Events and Their Names, Hackett Publishing Company, 1988.
[12] M. Brand, The language of not doing, Amer. Philos. Quarterly 8 (1) (1971).
[13] M. Bratman, Intentions, Plans, and Practical Reason, Harvard University Press, Cambridge, MA, 1987.
[14] H.-N. Castaneda, Intensionality and identity in human action and philosophical method, Nous 13 (1979) 235–259.
[15] B.F. Chellas, Modal Logic: An Introduction, Cambridge University Press, Cambridge, UK, 1980.
[16] P.W. Cheng, L.R. Novick, Causes versus enabling conditions, Cognition 40 (1991) 83–120.
[17] P. Cohen, H. Levesque, Intention is choice with commitment, Artificial Intelligence 42 (1990) 213–261.
[18] P. Cohen, J. Morgan, M. Pollack (Eds.), Intentions in Communication, MIT Press, Cambridge, MA, 1990.
[19] D. Davidson, Actions and Events, Clarendon Press, Oxford, 1989.
[20] D. Davidson, Actions, reasons, and causes, in: Actions and Events, Clarendon Press, Oxford, 1989, pp. 3–20. Originally published in 1963.
[21] D. Davidson, Intending, in: Actions and Events, Clarendon Press, Oxford, 1989, pp. 83–102. Originally published in 1978.
[22] D. Davidson, The logical form of action sentences, in: Actions and Events, Clarendon Press, Oxford, 1989, pp. 105–148. Originally published in: N. Rescher (Ed.), The Logic of Decision and Action, University of Pittsburg Press, 1967.
[23] E. Davis, Representations of Commonsense Knowledge, Morgan Kaufmann, San Mateo, CA, 1990.
[24] D.R. Dowty, Word Meaning and Montague Grammar, Kluwer Academic, Dordrecht, 1991.
[25] T. Eiter, G. Gottlob, On the complexity of propositional knowledge base revision, updates, and counterfactuals, Artificial Intelligence 57 (1992) 227–270.
[26] J.A. Fodor, Three reasons for not deriving 'kill' from 'cause to die', Linguistic Inquiry 1 (4) (1970).
[27] L.T.F. Gamut, Logic, Language, and Meaning, Vol. 2: Intensional Logic and Logical Grammar, University of Chicago Press, Chicago, IL, 1991.
[28] M.L. Ginsberg, Counterfactuals, Artificial Intelligence 30 (1986) 35–79.
[29] A. Goldman, A Theory of Human Action, Princeton University Press, 1970.
[30] A. Goldman, Action, causation, and unity, Nous 13 (1979) 261–270. This is a reply to Castaneda's paper in the same issue.
[31] B.J. Grosz, The contexts of collaboration, in: Proc. 5th International Colloquium on Cognitive Science, 1998.
[32] J.Y. Halpern, Y. Moses, A guide to the modal logics of knowledge and belief: A preliminary report, Technical Report, IBM Research Laboratory, 1985.
[33] G. Harman, Change in View, MIT Press, Cambridge, MA, 1986.
[34] H.L.A Hart, T. Honoré, Causation in the Law, 2nd ed., Clarendon Press, Oxford, 1985.
[35] P.J. Hayes, Short time periods, in: Proc. IJCAI-87, Milan, Italy, 1987, pp. 981–983.
[36] P. Horwich, Asymmetries in Time: Problems in the Philosophy of Science, MIT Press, Cambridge, MA, 1987.

[37] D. Israel, J. Perry, S. Tutiya, Actions and movements, in: Proc. IJCAI-91, Sydney, Australia, 1991, pp. 1060–1065.
[38] R. Jackendoff, Semantic Structures, MIT Press, Cambridge, MA, 1991.
[39] H. Katsuno, A. Mendelzon, On the Difference between Updating a Knowledge Base and Revising It, Cambridge Press, Cambridge, UK, 1992, pp. 183–203. Long version with proofs.
[40] J. Kim, Causes and counterfactuals, J. Philos. LXX (17) (1973) 570–572.
[41] J. Kim, Noncausal connections, Nous 8 (1974) 41–52.
[42] K. Konolige, Belief and incompleteness, in: Formal Theories of the Commonsense World, Ablex Publishing Corporation, Norwood, NJ, 1985, pp. 359–404.
[43] S.A. Kripke, Naming and Necessity, Harvard University Press, Cambridge, MA, 1980.
[44] S.C. Levinson, Pragmatics, Cambridge University Press, Cambridge, UK, 1983.
[45] D. Lewis, Counterfactuals, Harvard University Press, Cambridge, MA, 1973.
[46] D. Lewis, Causal explanation, in: Philosophical Papers, Oxford University Press, Oxford, 1986, pp. 214–240.
[47] D. Lewis, Events, in: Philosophical Papers, Oxford University Press, Oxford, 1986, pp. 241–269.
[48] J.L. Mackie, The Cement of the Universe, A Study of Causation, Oxford University Press, Oxford, 1988.
[49] J. McCarthy, Notes on formalizing context, in: Proc. IJCAI-93, Chambéry, France, 1993, pp. 555–560.
[50] D. McDermott, A temporal logic for reasoning about processes and plans, Cognitive Sci. 6 (1982) 101–155.
[51] M. Moens, M. Steedman, Temporal ontology and temporal reference, Comput. Linguistics 14 (2) (1988) 15–28.
[52] R.C. Moore, A formal theory of knowledge and action, in: Formal Theories of the Commonsense World, Ablex Publishing Corporation, Norwood, NJ, 1985.
[53] L. Morgenstern, Foundations of a logic of knowledge, action, and communication, Ph.D. Dissertation, New York University, 1988.
[54] G.M.P. O'Hare, N.R. Jennings (Eds.), Foundations of Distributed Artificial Intelligence, Wiley, New York, 1996, pp. 169–186.
[55] C.L. Ortiz, The semantics of event prevention, in: Proc. AAAI-93, Washington, DC, 1993, pp. 683–688.
[56] C.L. Ortiz, Causal pathways of rational action, in: Proc. AAAI-94, Seatle, WA, 1994, pp. 1061–1066.
[57] C.L. Ortiz, Worlds of change: Counterfactual reasoning and causation, Ph.D. Dissertation, University of Pennsylvania, Department of Computer and Information Science, 1996.
[58] C.L. Ortiz, Explanatory update theory: Applications of counterfactual reasoning to causation, Artificial Intelligence 108 (1999) 125–178.
[59] M. Pollack, Inferring domain plans in question-answering, Ph.D. Dissertation, University of Pennsylvania, 1986.
[60] C. Rieger, An organization of knowledge for problem solving and language comprehension, Artificial Intelligence 7 (1976) 89–127.
[61] R.C. Schank, R.P. Abelson, Scripts, Plans, Goals, and Understanding, Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.
[62] Y. Shoham, Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence, MIT Press, Cambridge, MA, 1988.
[63] Y. Shoham, Varieties of context, in: V. Lifschitz (Ed.), Artificial Intelligence and Mathematical Theory of Computation; Papers in Honor of John McCarthy, Academic Press, New York, 1991, pp. 292–408.
[64] E. Sosa, Causation and Conditionals, Oxford University Press, Oxford, 1975.
[65] L.A. Stein, L. Morgenstern, Motivated action theory: A formal theory of causal reasoning, Artificial Intelligence 71 (1994) 1–42.
[66] L. Talmy, Semantic causative types, in: Syntax and Semantics, Vol. 6, Academic Press, New York, 1976.
[67] L. Talmy, Force dynamics in language and cognition, Cognitive Sci. 12 (1988) 49–100.
[68] J.J. Thomson, Acts and Other Events, Cornell University Press, Ithaca, NY, 1977.
[69] T. Yagisawa, Counterfactual analysis of causation and Kim's examples, Analysis 39 (2) (1979).
[70] Z. Vendler, Causal relations, J. Philos. LXIV (21) (1967) 704–713.
[71] Z. Vendler, Facts and events, in: Linguistics in Philosophy, Cornell University Press, Ithaca, NY, 1967, pp. 122–146.
[72] B. Vermazen, Negative acts, in: B. Vermazen, M. Hintikka (Eds.), Essays on Davidson, Clarendon Press, Oxford, 1985. See also Davidson's reply to Vermazen in the same volume.

[73] G.H. von Wright, On the logic and epistemology of the causal relation, in: E. Sosa (Ed.), Causation and Conditionals, Oxford University Press, Oxford, 1975, pp. 105–124.
[74] D.S. Weld, J. de Kleer (Eds.), Readings in Qualitative Reasoning about Physical Systems, Morgan Kaufmann, San Mateo, CA, 1990.
[75] M. Winslett, Reasoning about actions using a possible models approach, in: Proc. AAAI-88, St. Paul, MN, 1988, pp. 89–93.