Causal Pathways of Rational Action

Charles L. Ortiz, Jr.*
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
clortiz@linc.cis.upenn.edu

Abstract

A proper characterization of a rational agent's actions involves much more than simply recounting the changes in the world affected by the agent. It should also include an explanatory account connecting the upshots of an agent's actions with the reasons behind those actions, where those upshots might represent actual changes (either intentional or unintentional) or merely counterfactual possibilities. The conventional view of action makes it difficult to distinguish, inter alia, cases of attempts, accidents, coercions, or failures — such distinctions useful to agents engaged in recognizing or assigning responsibility for actions. Such a view also makes the characterization of actions that do not involve physical change, such as maintenance events, difficult, as well as the proper representation of negative actions; the latter commonly appearing in explanations and as objects of an agent's intentions. In this paper, I present a formal analysis of these sorts of actions in terms of the causal pathways joining an agent's intentions with his actions.

Introduction

Consider a simple situation in which the following action is observed:

(1) John tried not to spill the coffee by holding the saucer steady but failed.

To capture conditions for the occurrence of instances of such an action by way of traditional representations that equate actions with pairs of world-states (McCarthy & Hayes 1969) would be difficult given that, first, the observable pre- and post-conditions of the above action could apply equally well to instances of intentionally spilling the coffee by tipping the saucer or to cases of unintentionally spilling the coffee while involved in some other activity[1]. Further, (1) also makes reference to an *attempt* at performing a *negative action* (not spilling the coffee) where the referenced negative action is to be related to a "positive" means action: the holding the saucer steady action. Though (Pollack 1986) and (Israel, Perry, & Tutiya 1991), for example, have examined how one action can be a means for the performance of another action, that work leaves open the question of how a positive (or negative action, for that matter) can be a means of not performing some ends action[2]. In addition, such a negative action must be represented in a fashion that precludes the possibility of identifying it with the non-occurrence of the *spilling of the coffee* action. Otherwise, it would be impossible to distinguish the referenced negative action with any other action that might have occurred. The action reported in (1) further suggests a deviation in the usual causal pathway joining the agent's prior intention (to not spill) with the final result. This is reported as a failure: failures representing another class of negative action, members of which can be characterized as unintentional. This last point suggests that negative actions cannot be simply identified with non-movement or non-change; in the example, the referenced failed action *did* result in change: the spilling of the coffee.

In addition, any successful analysis of (1) should uncover a number of implicit events: in the case of successful performance on the part of the agent would be the presence of a *maintenance* action involving no change with respect to the fact that the coffee is in the cup. It does not seem possible to reconcile the traditional view of action (under which nothing has "happened") with the intuition that here the agent has, in fact, *done* something. Additionally, (1) would normally also be re-describable as an instance of an *accident*. That an accident cannot simply be equated with an unintentional action is plain enough: my going to a lecture might be intentional but the fact that by doing so I thereby also occupy the last available chair is incidental: it represents a *side-effect* of my intentions

*This research was supported by the following grants: ARO no. DAAL 03-89-C-0031 Prime and DARPA no. N00014-90-J1863.

[1]Though (McDermott 1982) presents an alternative representation of events by which one can, for example, represent actions such as "running around the track," that work does not address the issues discussed here of generation relations, particularly involving negative actions, maintenance actions, or attempts, failures and accidents.

[2]For a discussion of negative events see (Brand 1971).

(Cohen & Levesque 1990).

Contemporary theories of rational behavior, such as the logic of belief, desire, and intention (BDI) of (Cohen & Levesque 1990)(C&L), have argued persuasively for the need for a notion of prior intention (or commitment) to action. However, the mere presence of an action does not presuppose the existence of a causally responsible intention; neither can the presence of an intention assure success. These are simply idealizations, as illustrated by example (1). In what follows I first present a brief overview of C&L in which I will frame the subsequent analysis. I then examine *means-end causal pathways* involving notions of generation and enablement as tied to an agent's intentions. I go on to examine maintenance actions as characteristic of actions involving no change. Finally, I examine cases of accidents, attempts, and failures as representative of deviations from means-end causal pathways. In this analysis, I take the agent's mental state as playing a crucial role in the proper characterization of such actions and argue that the characterization of these sorts of actions should involve counterfactuals.

Representation

C&L's logic models belief with a weak S5 modal logic where possible worlds are linear sequences of *basic* event types (indexed by the integers); satisfaction is then defined relative to a model, world-time pair, and a variable valuation. An agent i's belief in some proposition ϕ is expressed in their language by statements of the form $bel(i, \phi)$, while an agent's goals (consistent desires) are captured by statements of the form $goal(i, \phi)$. The semantics of *Bel* and *Goal* are both given in terms of possible worlds. An agent's intentions, written $intend(i, \alpha)$, are composite objects modeled as *persistent goals*: these are goals which agent i will maintain until α is achieved or until i believes α is no longer possible. Intentions have the further important property that they are not closed under logical consequence. Their model also makes use of the following modal temporal operators: $\Diamond\phi$ means ϕ is eventually true in the current world (which includes the current moment), $\Box\phi =_{def} \neg\Diamond\neg\phi$, and $later(\phi) =_{def} \neg\phi \wedge \Diamond\phi$. The modal operators $happens(\alpha)$ and $done(\beta)$ refer to the actions α and β as, respectively, happening next or as having just happened in the current world-time point (with an optional extra argument standing for the agent of the action). Complex action descriptions are possible by way of statements in dynamic logic where the set of actions is closed under nondeterministic choice $(\alpha|\beta)$, sequencing $(\alpha;\beta)$, tests $(p?)$, and iteration (α^*). The reader is referred to (Cohen & Levesque 1990) for details on the logic.

In order to refer to future *possibility*, I will add the following branching modal operators to C&L. $\Diamond_B\phi$ means that among all of the possible worlds with pasts (not including the current moment) identical to the real one, there is one in which ϕ holds. $\Box_B$ is defined as usual as $\neg\Diamond_B\neg\phi$. In order to model concurrent actions I will introduce the function $+$: $\alpha + \beta$ names the action consisting of the simultaneous occurrence of α and β. Possible worlds stand, as before, for sequences of events, however, each event is now taken from the lattice structure built up out of $+$ and the primitive NIL standing for inaction. I also define: $\phi > \psi$ as $\Box_B[\phi \supset \psi]$. This says that the material conditional holds in all possible futures. Where the conditional is evaluated at some time in the past and ϕ does not hold in the real world, then $>$ stands for counterfactual dependence. There are well known problems with this idealization which I will later discuss. To capture some notion of bringing about I define: $happens(e \rightsquigarrow \phi) =_{def} happens(\neg\phi?; e; \phi?) \wedge \forall e'.e' \leq e \supset happens(e'; \neg\phi?)$, that is, e leads to ϕ just in case it results in ϕ and any subsequence (specified by the ordering $\leq$) results in $\neg\phi$. In order to formalize a notion of generation I will introduce the following variant of *happens*: $happens(i, t, \alpha) =_{def} \exists t'.happens(i, t'?; \alpha; t' + t?)$ which states that t represents the duration of α where, in C&L, if t is an integer then $t?$ specifies the time as t; similarly for $done(i, t, \alpha)$. Finally, $not(\alpha)$ will stand for any instance in which: $\models happens(not(\alpha)) \equiv \neg happens(\alpha))$ and $basic(\alpha)$ will be true just in case ϕ is primitive. I will also assume that this modified version is extended to allow singular terms for actions where each action term is grounded in some basic event sequence. This is straightforward.

Means-end Causal Pathways

The notion of one action representing a *way of* performing another seems central to means-end reasoning (Pollack 1986; Israel, Perry, & Tutiya 1991): this relation is often referred to as *generation*, after Goldman's treatment (Goldman 1970). So too does the notion of enablement: that is, one action performed in order to perform another (Pollack 1986; Balkanski 1993; Di Eugenio 1993). Examples of generation are often reported by way of the *by* locution: *He signalled by waving* or *He turned on the light by flipping the switch.* Goldman notes that the generation relation is irreflexive, anti-symmetric, and transitive. In Goldman's theory, actions are triples of act-type, agent, and time where each action is either generated by some other action or represents a basic action or movement. Pollack formalized Goldman's notion of generation via a notion of *conditional generation*. Her formalization essentially associates a condition, c, with the occurrence of two distinct act-types, α and β, such that α generates β just in case α and β both occur at the same time, it is always the case that if c holds and α occurs then β occurs, and neither α nor c are sufficient for β to occur. For the case of turning on a light by flipping a switch, c might affirm that the light and switch are connected. Unfortunately, this approach cannot be used to explain instances such as (1) nor the following:

(2) By not lowering my hand I signalled again to the auctioneer.

Here, whatever c is chosen as the generating condition for the not-lowering action could just as easily be used to incorrectly conclude that some other non-occurring action generated the signalling.

Another difficulty with Pollack's formalization is that, given the potentially infinite number of qualifications to any action, there is no c such that α will always generate β under c: one can always find an exception. Further, her definition involves a second order statement. In this paper, I suggest that generation be analyzed counterfactually: in short, if α had not occurred β wouldn't have either. Such a definition correctly handles (2). This was, in fact, one of the clauses in Goldman's original definition[3], but was questioned by Pollack. The example Goldman used was the following. If some agent extended his arm in order to signal then we can say that if he had not extended his arm, he would not have signalled. Pollack observes that if the agent also intended to signal then there is no reason to suppose that the intention to signal would not survive the counterfactual supposition, particularly if there was some other means of signalling available to the agent. In other cases, however, the intention would not survive, else one would be forced to deny the validity of reasonable counterfactuals such as the following:

(3) If Oswald had not shot Kennedy, then Kennedy would be alive today.

This counterfactual holds because we have good reason to believe that, at that moment, a shooting was the only means reasonably available to Oswald; in such a case, the intention to kill would not survive the counterfactual supposition. One possible solution to the objection raised by Pollack is to introduce particulars for act-tokens by way of statements of the form: $happens(e) \wedge type(e, wave)$ and argue that if *that* arm extension had not occurred then *that* signal would also have not occurred. An alternative, and the one taken here, is to argue that the counterfactual dependence holds between β and both the intention to β and α: to make an analogy with causation, the intention to β and α are on the same causal pathway to β.

A case of α generating β can then be defined as [4]:

$$\begin{aligned} gen1(\alpha,\beta) \equiv\ & \alpha \neq \beta \\ & \wedge [happens(i,t,\alpha) > happens(i,t,\beta)] \\ & \wedge [(\neg intcnd(i,\beta) \\ & \quad \wedge \neg happens(i,t,\alpha)) > \neg happens(i,t,\beta)] \end{aligned} \tag{1}$$

[3] Goldman included this clause in order to handle *branching acts*. See (Goldman 1970). This seems reason enough for its inclusion.

[4] All axioms that appear in this paper are assumed to hold for every model and world-time pair. All unbound variables are further assumed to be universally quantified.

$$\begin{aligned} gen(\alpha,\beta) \equiv\ & gen1(\alpha,\beta) \\ & \vee [\exists\gamma.gen1(\alpha,\gamma) \wedge gen(\gamma,\beta)] \end{aligned} \tag{2}$$

This inductive definition is needed because counterfactual dependence is not generally transitive (Ginsberg 1986). The first axiom states that α and β must be distinct (this enforces ireflexivity) and that β counterfactually depends on both α and $intend(i,\beta)$. The second axiom simply states that two actions are related by generation just in case there is a chain of counterfactual dependencies between the pair. This approach depends on a body of basic *generation knowledge* of the form: $happens(i,t,\gamma) \wedge c \supset happens(i,t,\delta)$, where these axioms can be separately qualified (Ginsberg & Smith 1988). It also depends on a more restrictive treatment of $>$ along the lines of (Ginsberg 1986) so that if, for example, the arm extension (α) had not occurred then one only considers possible worlds that *must* follow from this revision and causal knowledge: alternative means of signally will not be considered since the causal factor (intention) has been retracted [5]. In order to ensure the proper direction of generation and its characteristic antisymmetry, one approach to explore would be to *unprotect* generation knowledge as in (Ginsberg 1986), i.e., allow it to be retracted so that there would exist possible worlds in which the agent had not β-ed but had α-ed.

One problem with the definition conjectured in 1 involves the following example of Goldman, *George jumps 6'. John outjumps George by jumping 6'3"*. Under the counterfactual analysis the best one case say is that John outjumped George by jumping over 6'; this seems reasonable. A more serious problem is how to ground basic negative actions; recall that "positive" actions were grounded in basic agent movements. Consider, for example, *refraining from going to school today*, where there is no more basic action, α, such that if α had not occurred the positive counterpart of the above would have. Further, one cannot simply equate not-α with $\neg happens(\alpha)$, as discussed earlier. I suggest that the notion of a basic action is a *dynamic notion* which varies according to an agent's mental state — as well as possibly some external agent's expectations. Actions are then grounded in these partial mental state descriptions: if the agent had wanted to go to school, he would have.

Given the above, the composite *by* action can now be defined as follows:

$$\begin{aligned} & happens(i,t,by(\alpha,\beta)) \equiv \\ & \quad happens(i,t,\alpha) \wedge happens(i,t,\beta) \wedge gen(\alpha,\beta) \end{aligned} \tag{3}$$

That is, the action $by(\alpha,\beta)$ is said to occur just in case both α and β occur over identical spans of time and, moreover, α generates β. An agent can now intend

[5] In addition, the essential temporal asymmetry (see (Lewis 1986)) of counterfactuals must be addressed: the nearest possible worlds should be those in which the past is kept as stable as possible.

to perform some β by performing some more basic α. Where the agent is successful, this represents the standard means-end causal pathway.

Turning now to enablement, it seems that central to this relation is the notion of "bringing about a possibility." Consider alternatives such as that suggested in (Balkanski 1993). On her analysis, α enables β just in case: (i) the time of α is prior to the time of β, and (ii) there is a set of conditions, C, such that one of the conditions in C, C_i, holds as a result of the performance of α, and either: there is a third action γ, and γ conditionally generates β under C, or C is the executability condition on β There is a serious problem with this definition, however. Consider the following simple situation involving filling a bucket with water. Suppose that there are two faucets and the bucket is currently positioned under one of them. According to Balkanski's definition, transporting the bucket to the other faucet enables filling the bucket, even though that action is *already* possible in the initial situation: this seems non-intuitive. However, if one instead stipulates that α must render β possible, one encounters the following difficulty.

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

where the enabled action is already possible. It should further be pointed out that enabled actions must be intentional: notice the unacceptability of, *stopping on the way to the gate enabled him to miss the train*, unless the agent was actively trying to miss the train. This appears to be a consequence of their central role in means-end reasoning.

The following definition overcomes these problems:

$$\begin{aligned} &happens(i,t,enables(e_2)) \equiv \qquad (4)\\ &\quad intends(i,e_2) \wedge \exists e_1.happens(i,t,e_1)\\ &\quad \wedge [happens(e_1) > \Diamond_B happens(e_1;e_2)]\\ &\quad \wedge [\neg happens(i,t,e_1) > \neg\Diamond_B happens(not(e_1);e_2)] \end{aligned}$$

The use of counterfactuals assures us that if e_1 had not occurred then e_2 would not have been immediately possible and it, therefore, happily also rules out those circumstances in which e_2 might have eventuated on its own. Example (4) is also handled properly: there is no restriction that e_2 be initially impossible. Once again, 4 requires a treatment of $>$ along the lines of (Ginsberg 1986; Winslett 1988) [6]. Further, the set of actions quantified over in 4 should be restricted to some subset or context of actions as in (Ortiz 1993). This would preclude the possibility of identifying aberrant events with the event *not-*e_1 in the evaluation of, say, example (4) so that when evaluating the second counterfactual, one does not consider quicker means of transport that would represent a departure from the norm.

[6] NB: Axiom 1 correctly assigns some e_1 (and *not enables*(e_2)) the role of generating action in cases where e_1 satisfies 4 if one uprotects axiom 4.

Non-movement actions

Maintenance events differ from accomplishments in that they do not involve any change with respect to the proposition being maintained: by repeatedly pushing a door (call each such component event an α) I can maintain the door in a closed position (call this condition ϕ), however, the door must have been in a closed position to start with[7]. If I initially closed the door, that action would have been distinct from my later maintaining action. In addition, the condition which is maintained, ϕ, must be counterfactually related to each component α: if I hadn't been pushing the door, it could have opened at some point, possibly because someone was pushing from the other side. This cannot be determined simply by observing the scene: I might simply be pushing a locked door. Furthermore, each α does not have the property that if it had not occurred ϕ would necessarily have come about. Consider a case in which I am maintaining coffee in a cup while walking on a ship that is rocking. Suppose further that whenever the cup "dips" over an angle greater than θ, the coffee spills. I might choose to "straighten" the cup as soon as it rotates no more than some $\theta - \delta$: such an action is nonetheless a component maintaining even though I could have just as well salvaged the coffee by waiting a bit longer.

There appears to be a close relationship between maintenance actions and preventions (Ortiz 1993). Whereas a prevention is a relation between a real event and a hypothetical event, a maintaining can best be viewed as a process composed of smaller events (α's), where each component α inhibits progress towards the bringing about of ϕ but does not render ϕ impossible: each push I initiate does not prevent the door from ever opening — only for the extent of that particular push. Further, if I lock the door, thereby preventing anyone from ever opening it, I am not maintaining the door closed. Notice that maintenance events share a property of all processes: they are not necessarily homogeneous (Shoham 1988) over the interval in which they occur: in the example, the pushes originating on the opposite side of the door might be intermittent. In addition, each step (α) need not be of the same type. For example, suppose I have a rope with a knot at the center which I am attempting to maintain within some spatial interval. Periodically, someone pulls on either side: if I feel a tug to the left, I pull to the right, and vice-versa. In each case, my action is of a distinct type: a left-pull versus a right-pull.

These properties can be captured by the following:

[7] (McDermott 1982) discusses an analogous notion of *protecting* a fact but leaves it as an open problem. (Di Eugenio 1993) discusses maintenance actions in the context of instructions.

$$done(i,t,m(\phi)) \equiv \quad (5)$$
$$\exists\alpha.done(i,t,\alpha) \wedge \Diamond_B later(\neg\phi)$$
$$\wedge\ \exists d \forall\beta[happens(\beta \rightsquigarrow \neg\phi) > f(\beta,\neg\phi) \geq d]$$
$$\wedge\ [\neg done(i,t,\alpha) >$$
$$\forall\gamma[happens(\gamma \rightsquigarrow \neg\phi) > f(\gamma,\neg\phi) < d]]$$

This says that some α generates a component maintaining, $m(\phi)$, if it inhibits progress towards ϕ; i.e., if all the possible futures (recall the definition for $>$) leading to $\neg\phi$, represented by the sequence β, have a cost greater than d; whereas if α had not occurred, all of the events, γ, leading to $\neg\phi$ would have had a lower cost. The function f is meant to capture a sense of "progress" towards ϕ (see below). The second clause in the definition further constrains the coming about of $\neg\phi$ to remain possible: this is necessary so that α not prevent $\neg\phi$ from ever coming about. A maintenance event can now be represented as a process or instance of $\phi?;[m(\phi)|(x;\phi?)]^*;m(\phi);[m(\phi)|(x;\phi?)]^*$, for some basic event, x; that is, as a possibly inhomogeneous sequence consisting of $m(\phi)$'s, with the further restriction that ϕ be true throughout.

By employing counterfactuals, the above definition correctly captures the tendency towards $\neg\phi$ that the agent is inhibiting. Further, since α can occur concurrently with some other event, the definition allows one to model situations such as those involving a simultaneous push and pull — *no* net movement — or for α to represent a reaction (e.g., tugging to the right when you feel a tug to the left). Finally, though the issue of progress towards $\neg\phi$ is obviously a difficult one to resolve it does appear necessary in order to explain cases such as the example involving maintaining coffee in a cup, where the coffee can still change its location within the cup. Such a function is analogous to a heuristic search function that might be employed by a planner. Many problem domains readily suggest natural progress functions: in trying to maintain someone from getting close to some object, a natural metric would be the distance towards the object; in trying to maintain a tower at a certain height, a natural metric would be the number of blocks stacked so far; and in trying to maintain a bucket under a faucet, a natural metric would be the distance from the edge of the faucet to the edge of the bucket.

Abnormal Causal Pathways

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 generate α. If i fails then either: (i) some expected circumstance necessary for the generation of 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[8]. For example, consider the following:

(5a) John tried to escape but was caught.

(5b) John tried to remove the stains with soap and water.

(5c) 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 (5)b, John might have perfect knowledge about 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 axiom seems to capture these intuitions:

$$happens(i,t,try(\alpha)) \equiv \neg basic(\alpha) \quad (6)$$
$$\wedge\ \exists\beta.happens(i,t,\beta) \wedge intend(i,by(\beta,\alpha))$$

This states that an agent i attempts α just in case i performs some β with the intention of performing α. By the definition for *by* and the fact that intentions in C&L are closed under logical equivalence, it follows that i believes (whether correctly or not) that β will generate α. The case of basic actions is much more problematic and here I simply assume that basic actions always succeed (See (Pollack 1986) for a discussion) [9].

The notion of a failure is now captured as follows:

$$happens(i,t,fail(\alpha)) \equiv \quad (7)$$
$$happens(i,t,try(\alpha)) \wedge \neg happens(i,\alpha)$$

"You can't fail if you've never tried," as they say. Notice that it is not necessary that α be physically possible to start with. A consequence of the above definition, together with axioms from C&L, are the following reasonable inferences: an agent will never try to fail and if an agent tries, that attempt (i.e., $try(\alpha)$) cannot fail (i.e., $fail(try(\alpha))$ is impossible).

Turning now to cases of accidents, consider what appears to be a reasonable default for rational agents:

$$\neg intend(i,\alpha) \Rightarrow \neg later(happens(i,\alpha)) \quad (8)$$

[8] (Pollack 1986) discusses similar issues in the context of plan recognition.

[9] One problem with this definition, which I will not address, stems from the fact that if an agent tries to α it must fully believe that the means action will generate α, whereas in actuality it might entertain only a partial belief.

that is, if an agent doesn't intend to perform some α then it normally will not. Notice the distinction between the absence of an intention in this case and an intention to not perform α. The latter involves, as discussed earlier, the commitment to perform some more basic action as a way of performing the negative action: such commitments playing an important role in the architecture of rational agents (Bratman, Israel, & Pollack 1988). Axiom 8 can be defeated in cases of an accident (an agent spilling coffee for example), or in cases of *coercions*, in which some agent forces another agent to perform some action against his will. The first case is possible because an agent's intentions are not closed under implication in C&L. That is, an agent does not intend all of the side-effects of its intentions. Given this property, an accident can be defined as follows.

$$\begin{aligned} &happens(i,t,accident(\alpha)) \equiv \\ &\quad happens(i,t,\alpha) \wedge \neg bel(i, happens(i,\alpha)) \\ &\quad \wedge [knows(i, happens(i,\alpha)) > \\ &\qquad happens(i, try(not(\alpha)))] \end{aligned} \tag{9}$$

In this case agent i performs some α without being aware (possibly as a side-effect of some other intended action). However, in the case of an accident it is also necessary that the agent would have tried to avoid α, if it could have more accurately predicted the future. Notice that accidents include cases of failures, but not necessarily vice versa[10].

Summary and Conclusions

In this paper I explored a broader view of action than is possible through traditional accounts that equate an action with simply the bringing about of a condition. In the process, I discussed a number of action types, such as negative actions and maintenance actions, that are best viewed as depending counterfactually on a more primitive means action as wells as partial a mental state description. This observation led to a generalization and simplification of previous accounts of two important means-end relations — generation and enablement. I argued that actions tied to an agent's intentions by way of these two relations characterized the normal causal pathway of action whereas cases of failures, accidents, and coercions exemplified deviations from this pathway. Such a conceptualization of action is important to the architecture of rational agents that can recognize actions in order to ascertain abilities or assign responsibility as well as produce explanatory accounts of behaviors.

[10] Cases of coercions appear to represent deviations from the perspective of the normal evolution of deliberations from desire to intention formation. They are characterized by a concurrent desire (caused by another agent) *not* to perform some action while simultaneously entertaining the opposite intention.

Acknowledgments

I would like to thank Mike Moore, Mark Steedman, Bonnie Webber, and Mike White for comments on an earlier draft of this paper.

References

Balkanski, C. T. 1993. *Actions, Beliefs, and Intentions in Multi-Action Utterances.* Ph.D. Dissertation, Harvard University.

Brand, M. 1971. The language of not doing. *American Philosophical Quarterly* 8(1).

Bratman, M. E.; Israel, D. J.; and Pollack, M. E. 1988. Plans and resourse-bounded practical reasoning. *Computational Intelligence* 4:349–355.

Cohen, P., and Levesque, H. 1990. Intention is choice with commitment. *Artificial Intelligence* 42:213–261.

Di Eugenio, B. 1993. *Understanding Natural Language Instructions: a Computational Approach to Purpose Clauses.* Ph.D. Dissertation, University of Pennsylvania.

Ginsberg, M. L., and Smith, D. E. 1988. Reasoning about action ii: The qualification problem. *Artificial Intelligence* 35:311–342.

Ginsberg, M. L. 1986. Counterfactuals. *Artificial Intelligence* 30:35–79.

Goldman, A. 1970. *A Theory of Human Action.* Princeton University Press.

Israel, D.; Perry, J.; and Tutiya, S. 1991. Actions and movements. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1060–1065.

Lewis, D. 1986. Counterfactual dependence and time's arrow. In *Philosophical Papers.* Oxford University Press. 32–66.

McCarthy, J., and Hayes, P. 1969. *Some Philosophical problems from the standpoint of artificial intelligence.* Edinburgh University Press. 463–502.

McDermott, D. 1982. A temporal logic for reasoning about processes and plans. *Cognitive Science* 6:101–155.

Ortiz, Jr., C. L. 1993. The semantics of event prevention. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, 683–688.

Pollack, M. 1986. *Infering Domain Plans in Question-Answering.* Ph.D. Dissertation, University of Pennsylvania.

Shoham, Y. 1988. *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence.* MIT Press.

Winslett, M. 1988. Reasoning about actions using a possible models approach. In *Proceedings of the National Conference on Artificial Intelligence.*