Defeasibly Successful Action

Richmond H. Thomason
Philosophy Deparment
University of Michigan
Ann Arbor, MI 48109-1003
rich@thomason.org

Date of this version: April 8, 1998

Abstract

I present reasons why the relation between an action and its consequences—including the consequences that are goals—should be treated as defeasible.

Effects as infallible consequences of action

With almost no exceptions, the theories of action that have been proposed by philosophers, logicians, and (most recently) by researchers in Artificial Intelligence subscribe to the assumption that effects are consequences of actions that in some sense are necessary.

"Philosophy of action" is a recognized specialty in contemporary philosophy, and the literature on action is fairly extensive: see, for instance, (Care & Landesman 1968; Goldman 1970; Hornsby 1980). The relation of actions to their effects is formulated most clearly in the more specialized literature on the logic of action; see (Belnap & Perloff 1988; Chellas 1992; Czelakowski 1996; Segerberg 1982).

The approaches differ in detail, but the basic ideas are similar to those of dynamic logic. Actions are operators on the state of an indeterministic world; an action like [GoHome-by-Tomorrow] is similar to a necessity operator, in that its effects must take place in all the possible histories in which the action is performed. I.e., in all these histories, the agent must be home by tomorrow.

The assumption of these theories, then, is that the connection of actions to their effects is one of causal necessity. From a general philosophical standpoint, the conclusion that the action-effect relation is one of causal necessity will follow as soon as it is granted that the relation is causal. Many philosophers have found the assumption that causes are necessary so plausible as to seem axiomatic.[1]

The issue also arises in connection with planning formalisms deriving from STRIPS. In these formalisms, axioms connecting an action to its effects are monotonic, like the following axiom giving the effects of stacking one block on another.

$$(1)\ [Holds(Clear(x), s) \wedge Holds(Clear(y), s)] \rightarrow Holds(On(x, y), result(stack(x, y), s))$$

There is no explicit operator for historical necessity in these formalisms, as there is in logics of action like that of (Belnap & Perloff 1988). Nevertheless, the necessity of the relation between actions and their effects can be seen at the metalinguisic level. In every model in which an action occurs, its effect must also occur. Or we can make the point by using

[1] Perhaps the best-known examination of causal necessity is Immanuel Kant's *Critique of Pure Reason*; see (Kant 1961, p. 125). But also see, for instance, (Edwards 1957, p. 215), "To suppose there are some events which have a cause and ground of their existence, that yet are not necessarily connected with their cause, is to suppose that they have a cause which is not their cause."

provability: axioms such as (1) enable us to prove that the effects will be achieved if the action occurs. It will simply be inconsistent to suppose that an action is performed without its conventional consequences ensuing. This point holds not only for the classical planning formalisms, but also for more sophisticated modifications that are in part motivated by the need to accommodate operators with indeterminate effects.[2] See the concluding section of this paper for further discussion and comparison.

Problems with infallible effects

I want to argue that formalisms that make the connection between actions and their effects necessary are inadequate in a number of (related) ways: (1) they are contrary to common sense; (2) they do not match the way we speak about actions and their effects in natural language; (3) they do not provide a good mechanism for monitoring and reasoning about the reliability of plans; and (4) they are not readily applicable in some planning domains.

1. This point is made most easily with respect to the action logics. According these theories, when an action is performed and a change occurs, the change is an effect of the action only if it the action could not be performed without the change occurring. According to such a theory, if I turn the key in my car ignition, for instance, and the car starts, I can't claim that I started my car (that my action caused the car to start) without also claiming that the car must have started given the occurrence of my action. Note that, although (with some cars, anyway) the probability of the car starting on the key's being turned is high, this probability is not 1. Also, the circumstances under which the effect fails are difficult to enumerate exhaustively—in fact, this example is often used to illustrate the qualification problem.

[2]See (Giunchiglia, Kartha, & Lifschitz 1997).

But it is not just a matter of high probability. Suppose that I drop a glass and the glass breaks. According to a necessary effects theory, I broke the glass in this case (i.e., what I did caused it to break) only if there is no possible history in which I did what I did (dropped the glass) and it did not break. Again, this conclusion is highly improbable from the standpoint of common sense. From the standpoint of the best physics we have, we can imagine a situation in which the breaking of the glass depends on quantum effects, whose probability is as close to r as we like, for any $r \in (0, 1)$. Although when the probability is low I may be able to excuse myself by saying it was a freak occurrence, it seems to me to be contrary to common sense for me to excuse myself by saying I didn't break it, because I might have dropped the glass without its breaking.

2. Natural languages typically have ways of associating conventional or normal consequences with an action, and of distinguishing the initiation and ongoing performance of an action from the successful achievement of these consequences. In English, there is no regular linguistic relationship between the adjectives expressing states that conventionally ensue from an action and the verbs for achieving these effects. One pattern is exhibited by 'open'. This word is both an adjective, as in

(2) 'The door is open'

and a verb, as in

(3) 'She opens the door.'

The (possibly fallible) performance of an action is encoded in English by the progressive aspect, which involves a form of the auxiliary verb 'be' and the suffix '-ing':

(4) 'She is opening the door.'

(5) 'She was opening the door'.

The successful achievement of the conventional effects of the action is encoded in the simple past, as in (6) as well as in the perfective aspect, as in (7) and (8), which involve a form of the auxiliary verb 'have' and a suffix '-ed' or '-en'.

(6) 'She opened the door.'

(7) 'She has opened the door'.

(8) 'She had opened the door'.

Note the difference between these forms. Sentences (4) and (5) do not entail that she ever opened the door, whereas sentences (6), (7), and (8) do have this entailment.

Providing a model theoretic interpretation of these constructions is one of the central problems of linguistic semantics; a classical source for work on this problem is (Dowty 1979). The difficulty is this: there is very good linguistic evidence for saying that part of the meaning of (6) is that Alice was the agent of a process that culminated in a state of the door's being open; and the simple past tense sentence (6) entails that the door came to have this state. But the past progressive forms, such as (5), do *not* entail that the action succeeded.

Dowty's theory of the linguistic phenomena actually prefigures some of the later ideas that emerged in AI, with his notion of an "inertial world," which essentially is a world in which the normal consequences of actions occur. There have been later attempts to use nonmonotonic logic to improve on Dowty's theory; see (Asher 1992). The theory that I will develop below is close in many ways to this linguistic work.

3. The ultimate purpose of a planning formalism is to relate goals to sequences of performable actions that will achieve these goals. When the matter is put this way, it is hard to find any realistic planning problems in which the relation between the action sequences and the goals is infallible, although in many cases the effects are reliable enough so that we can discount their fallibility for many practical purposes. But AI planning formalisms do not provide any natural way to represent this fallibility or to take it into account.

Suppose, for instance that I have the goal of being in Boston, AT(I,BOSTON). I can form a one-step plan ⟨GOTO(BOSTON)⟩ for this purpose. But this operator (at least, in its abstract, undecomposed form) is not immediately performable. As soon as I put the action in an executable form (let's assume that I am at the proper airport gate, and that ⟨BOARD(PLANE)⟩ is such a form), it is perfectly possible to execute the operator without achieving the original goal; flights don't always arrive at their destinations.

Some forms of contingency planning or risk management involve taking the fallibility of actions into account. It can be perfectly reasonable to board a plane in order to get to Boston, and at the same time to warn the person who is to meet you what to do in case the flight is cancelled, or even to buy flight insurance.

I don't claim to have any way to formalize this sort of reasoning, but I suspect that it might be easier to do in a formalism that does not make it inconsistent to suppose than an action can occur without achieving its normal effects.

4. It is fairly usual, in formalizing the interpretation and generation of discourse, to think of speech acts as planning operators; the idea goes back at least as far as (Cohen & Perrault 1979).

But the goals of many typical speech acts are to some extent cooperative. This is true especially of speech acts that form half of a pair, actions like questions and proposals. The goals of these acts can't be achieved in one conversational turn, and it is not unusual for them to fail; in fact, there are conventional ways to decline to answer a question ('I don't know') or to

refuse a proposal ('Sorry, I'm busy'). The goal of a question is to obtain information, to rule out some epistemic alternatives that are open before the question is answered. The goal of a request is to obtain agreement on a plan, to rule out some practical alternatives that are open before the request is agreed on.

You have to think of a request as a hopeful bid for a course of action, a bid whose success depends on cooperation that may not be forthcoming. In a way, this is not too different from turning a key in the ignition, which is a sort of hopeful bid for a started engine, and which requires the cooperation not of an agent, but of nature. But this aspect can be safely ignored in formally many planning activities in which an agent is interacting with nature or artifacts, whereas I suspect that it can't be ignored in cases of social interaction.

If we try to formalize questioning and proposing as conventional planning operators, we are forced either to distort the effects or to add hidden preconditions. We could say, for instance, that the effect of requesting is to make it known that you have a certain desire. Or we could make a disposition to cooperate with the request a precondition of the request. On the former alternative, it will be difficult to relate the desired goal—securing a course of action—to the speech act of requesting, unless one assumes an implausibly optimistic theory of social interaction. On the second alternative, it will often be uncertain whether the preconditions of the action are met. Planning to meet these preconditions would produce highly artificial discourses like this.

(9) 'Would you open the door if I requested it?'
'Well, yes.'
'Please open the door.'

On the other hand, it would be incorrect to formalize requesting so that acceptance is secured in all situations that are not abnormal, or to formalize questioning so that an answer is secured in all situations that are not abnormal. The success of these actions depends in part on certain conditions that have to be inferred by agent modeling.[3] But even if these conditions are met, a question or a request can be frustrated for reasons that are difficult or impossible to enumerate explicitly; and to formalize this, a normality condition or the equivalent in some other nonmonotonic formalism is needed in order to reason from the desired goal to the action.

Risky actions and utilities

In general, whether it is reasonable to perform an action that we know may not succeed depends on both the probability of success and the penalties associated with failure. Because of this, reasoning about fallible actions is clearly related to utilities.

I had originally intended to present the details of a result relating nonmonotonic reasoning to dominance reasoning, but the time for preparing an adequate presentation was too short, and the space it would require is too great. Instead, I will provide an informal presentation of the ideas here. And I will deposit a detailed document, (Thomason 1998), at the following internet location.

www.pitt.edu/~thomason/nm-dominance.html

A plan dominates another when its outcomes would be better regardless of circumstances.

[3] Is the person I am addressing disposed to be cooperative? Is he in a position to know the answer to this question? Is the course of action I am recommending likely to be unwelcome?

> **Example 1:** Suppose that I wake up on a snowy morning. My university may be closed because of snow, or may not be closed; I don't know which. I can stay home or go to work. I do not mind walking the short distance to my office in the snow; that isn't a factor. I reason that if I go to work and the university is open, that is better than staying home because I should not miss my classes. And if I go to work and the university is closed that is better, since I will get more done in a quiet day at the office than at home. This provides a dominance argument for going to work. The argument is qualitative, in that it is independent of the probabilities of the circumstances and of a numerical representation of the utilities.

On the other hand, dominance arguments assume that the plans and circumstances are independent.

> **Example 2:** Suppose that age 20 I reflect that either I can work hard or not, and that either I will be wealthy or poor. If I'm going to be wealthy it is better not to work. And if I am not going to be poor it is also better not to work. The dominance argument for not working has the same form as the preceding argument for going to the office, but it is invalid because whether I am wealthy or poor is not independent of whether I work hand.

In (Thomason & Horty 1997), a model theory is developed that provides for dominance reasoning about actions whose outcomes are uncertain. We use branching time models. At each moment of time, there is a global state. There are many agents, who can perform actions. The pattern of actions performed at a time and the global states constrain, but do not determine the subsequent global states. It is this that makes branching possible: the successor moments will represent the various global states that can ensue given the pattern of actions.

A history is a path through successive states. Agents have a preference relation over these histories; this relation is a transitive, antireflexive order.

By holding constant the actions of others, it is possible to define conditionals in these models whose antecedents are actions. For an agent, a conditional 'If I were to do A then P' holds, relative to a contemplated history h, in case P holds in all histories h' in which I do A and others do exactly what they do in h.

Given this apparatus, it is possible to define a notion of dominance over actions: A dominates B in case for every history h, every history that would ensue if I did A supposing h is no worse than any history that would ensue if I did B supposing h. These notions can be extended to sequences of actions. The main result of (Thomason & Horty 1997) is the soundness of this notion of dominance with respect to quantitative dominance in causal decision theory, (Gibbard & Harper 1978).

The result shows that the dominance reasoning characterized in the qualitative theory is in some sense correct. But it uses too strict a criterion to be useful in many practical cases. Suppose, for instance, that in Example 1 I take into account the fact that there is some risk in going to work; I may be run over while crossing the street. No matter how unlikely this alternative is, it creates an outcome in which staying home is preferable to going to work and destroys the dominance argument. This remains true even if we make the risks symmetrical, by adding an alternative in which an airplane crashes into my house.

Intuitively, it seems that common sense dominance arguments depend not only on independence considerations, but on ignoring risks that in some sense can considered to be negligable.

Making the connection between actions and their normal effects defeasible provides a way to achieve this, while maintaining the qualitative spirit of (Thomason & Horty 1997). We extend the action model by adding a preference relation over transitions; crudely, some transitions are normal and others are not. The idea, of course, is that transitions in which actions do not achieve their conventional effects are abnormal.

We redefine dominance by ignoring abnormal transitions. Now, A dominates B in case for every normal history h, every normal history that would ensue if I did A supposing h is no worse than any normal history that would ensue if I did B supposing h.

The main technical result of (Thomason 1998) extends the soundness result of (Thomason & Horty 1997) to a version of causal decision theory with nonstandard probabilities, in which abnormal transitions get infinitesimal probability. This shows that it is rational in some (perhaps arcane) sense[4] to neglect the fallibility of actions, in cases where the probability of abnormal outcomes is negligable with respect to that of normal outcomes.

Conclusion

The main purpose of this paper has been to interest the planning formalisms community in rethinking the relations between actions and their effects. I do not intend to disparage the work that is based on the assumption that effects are necessary consequences of causes; it is appropriate in many planning domains, and it has enabled the development of insights into the formalization and implementation of planning systems. But the arguments against the assumption are very persuasive, and I suspect that we will not be able to understand common sense practical reasoning in many everyday domains without exploring how to give the assumption up.

I think of work on indeterminate effects, as in (Giunchiglia, Kartha, & Lifschitz 1997; Lin 1996), as complementary to the problems that I am raising here. This work addresses a related but different problem; how to deal with cases in which the side-effects of an action are indeterminate. As far as I know, it has not been used to address the problem of how to formalize actions whose goals are indeterminate, in that (in abnormal cases, at least) the action may be performed while the goal is not achieved. However, the techniques that have been developed for dealing with indeterminate effects may well apply to formalizing reasoning with fallible actions.

In this note, I have recommended replacing effect axioms for actions, like (1), with nonmonotonic axioms such as the following, in which the effect is only achieved when circumstances are normal.

$$(10)\ [Holds(Clear(x), s) \wedge Holds(Clear(y), s) \wedge \neg ab(x, y, s)] \rightarrow Holds(On(x, y), result(stack(x, y), s))$$

I sketched how a theory of nondeterministic actions could be improved by representing the normal outcomes, and how these normal outcomes could be used to provide a more robust account of dominance relations between the outcomes of various action sequences.

I have not, however, begun to address the problem of how to provide a general extension of plan-based reasoning to the case in which actions may fail to achieve their conventional goals. For instance, I have not yet begun to think through the interactions of the use of normality in (10) with the uses of normality in frame-based and causal reasoning. There are a number of complex issues here that have to be resolved before the shape of a theory that does justice to the considerations mentioned here is at all clear. But I hope that this sketch will

[4] Causal decision theory with infinitesimal probabilities is a departure from the norm in two important respects.

at least convince some people that these issues are worth pursuing.

References

Asher, N. 1992. A default, truth conditional semantics for the progressive. *Linguistics and Philosophy* 15:469–508.

Belnap, N. D., and Perloff, M. 1988. Seeing to it that: a canonical form for agentives. *Theoria* 54:175–199.

Care, N. S., and Landesman, C., eds. 1968. *Readings in the Theory of Action*. Bloomington, Indiana: Indiana University Press.

Chellas, B. 1992. Time and modality in the logic of agency. *Studia Logica* 51:485–517.

Cohen, P. R., and Perrault, C. R. 1979. Elements of a plan-based theory of speech acts. *Cognitive Science* 3:177–212.

Czelakowski, J. 1996. Elements of formal action theory. In Fuhrmann, A., and Rott, H., eds., *Logic, Action, and Information: Essays on Logic in Philosophy and Artificial Intelligence*. Berlin: Walter de Gruyter. 3–62.

Dowty, D. R. 1979. *Word Meaning in Montague Grammar*. Dordrecht, Holland: D. Reidel Publishing Co.

Edwards, J. 1957. *Freedom of the Will*. New Haven, CT: Yale University Press. Originally published in 1754.

Gibbard, A. F., and Harper, W. L. 1978. Counterfactuals and two kinds of expected utility. In Hooker, C.; Leach, J.; and McClennen, E., eds., *Foundations and Applications of Decision Theory*. Dordrecht: D. Reidel Publishing Co. 125–162.

Giunchiglia, E.; Kartha, G. N.; and Lifschitz, V. 1997. Representing action: Indeterminacy and ramifications. *Artificial Intelligence* 95(2):409–438.

Goldman, A. I. 1970. *A Theory of Human Action*. Princeton, New Jersey: Princeton University Press.

Hornsby, J. 1980. *Actions*. London: Routledge and Kegan Paul.

Kant, I. 1961. *Critique of Pure Reason*. New York: St. Martin's Press. First published (in German), 1781. Translated into English by Norman Kemp Smith.

Lin, F. 1996. Abstract operators, indeterminate effects, and the magic predicate. In Buvač, S., and Costello, T., eds., *Working Papers: Common Sense '96*, 96–103. Stanford University: Computer Science Department, Stanford University. Consult http://www-formal.Stanford.edu/tjc/96FCS.

Segerberg, K. 1982. The logic of deliberate action. *Journal of Philosophical Logic* 11(2):233–254.

Thomason, R. H., and Horty, J. F. 1997. Nondeterministic action and dominance: Foundations for planning and qualitative decision. Unpublished manuscript. Available at www.pitt.edu/~thomason/dominance.html. (A version of this paper was published in Yoav Shoham, ed., *Theoretical Aspects of Rationality and Knowledge: Proceedings of the Sixth Conference (TARK 1996), Morgan Kaufmann, 1996.*

Thomason, R. H. 1998. Fallible action and normal dominance. Unpublished manuscript. Available at www.pitt.edu/~thomason/nm-dominance.html.