Ability and Action

Richmond H. Thomason
Artificial Intelligence Laboratory
University of Michigan
130 ATL Building
1101 Beal Ave.
Ann Arbor, MI 48109-2110
U.S.A.

rich@thomason.org

Abstract

This is part of a larger project that is motivated in part by linguistic considerations and by the philosophical literature in action theory and the logic of ability, but that is also meant to suggest ways in which planning formalisms could be modified to provide an account of the role of ability in planning and practical reasoning.
In this version for the CSR workshop, I have suppressed most of the linguistic and philosophical issues, concentrating instead on the reasoning and its formalization.

Introduction

The literature on ability is scattered. I find it useful to divide it into the following five areas.

Philosophy of action. Some of these philosophers (especially Kenny) were heavily influenced by Aristotle. Others (such as Richard Taylor) were working in a broader analytic tradition; most of these were concerned in one way or another with foundations of ethics, and especially with the free will problem.

Ordinary language philosophy. The literature in the narrow ordinary language and Oxford traditions is centered around (Austin 1956).

Analytic philosophy concerned with branching time, possible worlds and deontic logic. This is not a unified tradition, although there are some coherent subgroups (like the group concerned with agency and the literature deriving from (Belnap, Jr. & Perloff 1988)). For my purposes, by far the most important paper in this literature is (Cross 1986), which in turn subscribes to the possible-worlds approach to conditionals, especially (Stalnaker 1968).

The linguistics and philosophy of modal constructions. (Kratzer 1977) is an influential paper in this tradition.

Work in planning formalisms. Ability is usually not addressed explicitly in the AI tradition, although intuitions about ability of course enter into the causal axioms for various actions. (Lin & Levesque 1998) and (Meyer, van der Hoek, & van Linder 1999) are exceptions.

The project that is described in this paper is separable from the philosophical work, and, although the work of philosophers like J.L. Austin and Anthony Kenny played a part in motivating my approach, in general the connections to the general philosophical literature are fairly remote. I have found that philosophical work can often provide a helpful orientation for many projects in formalizing commonsense reasoning, but that it seldom provides really specific guidance. On the other hand, I believe that a sensitivity to linguistic ideas—particularly to the techniques and theories used in semantics—is an important part of the methodology of formalizing commonsense reasoning. Our commonsense ideas are encapsulated in our language, and we need to think clearly and rigorously about the linguistic evidence. Also—as in many cases in logical AI—the relevant literature in philosophical logic (as opposed to garden-variety, nonlogical philosophy) can be a very useful starting point. I begin this exercise by saying something about the language of ability, and the logical theories of ability that have been presented in the philosophical literature.

Ambiguity Issues

Apparently, 'can' is ambiguous, as well as indexical or context-sensitive; the ambiguities may be manifold, and multiple dimensions of context may be involved. Any serious study of how to formalize ability has to begin by sorting out the ambiguities.

I believe that 'can' is ambiguous in several ways, and that only one sense is especially relevant for practical reasoning. As briefly as possible, I will try to indicate the readings of 'can' that aren't relevant.

Possibility versus ability

(1a) My shoes can be under the bed.
(1b) It can be true that my shoes are under the bed.
(1c) ?My shoes are able to be under the bed.

(2a) I can prove that theorem.
(2b) ?It can be true that I will prove that theorem.
(2c) I'm able to prove that theorem.

The fact that Example (1b) is not unnatural, and is a paraphrase of Example (1a), and the fact that Example (1c) is

unnatural, are indicators of a usage of 'can' to indicate that a possibility is not excluded. Contrast this with Example (2a–c), which is the usage I will be concerned with.

Generic versus occasional

(3) I can lift that rock.
(4) I can lift a 50 pound rock.

Example (4) is *generic*, it attributes a property to an agent that holds under a wide variety of times and circumstances—perhaps to all that are "normal" in some sense. I'm not concerned directly with generic uses of 'can', though I assume that to the extent that the meaning of a "generic tense" can be predicted from the meanings of the corresponding generic sentences, the following account may illuminate generic uses of 'can' as well. The occasional sense, which would be the most natural way to understand Example (3), is the one that I am concerned with here.

Generic abilities are not sufficient for planning purposes. I have a generic ability to climb trees, but of course I can't climb *any* tree. A three inch lodgepole pine with its lowest branches 30 feet above the ground is beyond my climbing abilities. Suppose that I'm out hiking and spot a grizzly a hundred yards away. Grizzlies are unpredictable, there is a potential emergency here, and I need a plan. I look around at the trees, I spot an alpine fir, and I say to myself "I can climb that tree." If the grizzly charges, this judgment has to ensure that I will get up the tree I selected. A 'can' that only guarantees that *I might* succeed in climbing it is not helpful here. I need to be sure that on this occasion, I *will* get up the tree if I try. A generic ability to climb trees is not what is wanted here. Even a generic ability to climb *this* tree is irrelevant, if the circumstances under which it can be expected to apply are not in place. For planning purposes, our judgments of ability have to aim at providing successful results when put into practice.

In cases in which failures are noncritical, we can relax our standards somewhat; but even here, a judgment that a trial might fail tends to undermine the plan. Practical 'can's require success.

Further ambiguities

There is a further ambiguity or variability in the meaning of 'can' having to do with the status of the outcome. We often say that we can't do **a**, meaning not that a trial will fail to produce a performance of **a** but that it will produce consequences that are forbidden or undesirable; this is the sense in which I might tell someone I can't meet them at 2 because I have another appointment at that time. This is a natural extension of 'can' for planning purposes, and I believe that it fits well into the theory that I will develop. But I will not have more to say about it, at least in this draft.

Possible worlds semantics for 'can'

(Cross 1985) is a good beginning in investigating the semantics of 'can'. Like possible worlds theories of the conditional, Cross' account exploits a function that, given a clause ϕ and an agent A, selects a set of possible worlds that is in some sense "close" to the actual world w. Intuitively, these are the worlds that provide appropriate test conditions for A's "performance" of ϕ, or rather (since ϕ expresses a proposition, rather than an action) for A's taking steps to bring about an outcome in which ϕ holds. Cross' semantic rule for $< A>$is this.

(Can1) $M \models_w <A>\phi$ if and only if $M \models_{w'} \phi$ for some $w' \in g(\phi, a, w)$.

$<A>\phi$ is true at a world w if and only if ϕ is true at some world w' in $g(\phi, \mathbf{A}, w)$. The function g selects the set of worlds that, relevant to the circumstances in w, would provide appropriate "test conditions" for, as Cross puts it,

(5) testing whether the truth of ϕ is within **A**'s abilities in w.

The function meets the following two conditions.

(SB) If $\{w\colon M \models_w \phi\} \subseteq \{w\colon M \models_w \psi\}$ then $g(\phi, \mathbf{A}, w) \subseteq g(\psi, \mathbf{A}, w)$.
(AC) If $M \models_w \phi$ then $w \in g(\phi, \mathbf{A}, i)$.

According to Cross, $<A>\phi$ is true at w if and only if for some $w' \in g(\phi, \mathbf{A}, w)$, ϕ is true in w', where $<A>\phi$ is proposed as an adequate formalization of a sentence involving the application of 'can' to a subject formalized by A and a clause formalized by ϕ. This makes 'A can' a relativized modal possibility operator that is closely related to the relational operator $\Diamond\!\!\rightarrow$ of (Lewis 1973), which David Lewis proposed as a formalization of conditional 'might' constructions.

The relativization solves some of the obvious problems of using a standard, nonrelativized possibility operator to formalize 'A can', such as Kenny's objection ((Kenny 1976b; 1976a) that $\boldsymbol{Can}_A[\phi \vee \psi]$ does not entail $\boldsymbol{Can}_A\phi \vee \boldsymbol{Can}_A\psi$.[1] Also, the idea that the meaning of 'A can' is associated with a hypothetical test in which **A** is given a fair chance, under normal circumstances, at an attempt to perform an appropriate action, is very appealing. However, Cross' proposal is unintuitive in some important respects.

The clausal argument of 'can'

Cross is working within a logical framework in which actions are not available. In many other cases (deontic logic, for example) the conflation of propositions and actions, even if it is unintuitive, does not prevent the development of sophisticated formalisms that illuminate the logical issues. The same may be true here. However, in a formalism that does provide for action, it would be more natural to construe $\boldsymbol{Can}_A$ as an operator on actions, and that is what I will do below, when I have adopted such a formalism.

Is 'can' a possibility operator?

To put it roughly, Cross' theory of the 'can' of ability is based on an equivalence between 'I can' and 'If I tried I might'. I reject this equivalence: it seems to me that 'If I tried I would' is a more intuitive conditional explication.

[1] For the moment, I'll use $\boldsymbol{Can}_A$ for an unformalized representation of the 'can' of ability.

At this point, I will skip some detailed discussion of this issue. See the working version of the longer project at (Thomason 2003) for details.

To sum the matter up, the evidence concerning the modal status of 'can' appears to be mixed, making it difficult to provide a theory that is unequivocally supported by the evidence. However, I believe that the following theory is well enough supported to be plausible, and that the apparent counterexamples can be explained away in a principled way.

A conditional theory of ability

The logical situation with respect to ability is, I think, somewhat similar to the one that prevails with conditionals. According to the *variably strict* theories ((Lewis 1973) is an example), a conditional ***If*** ϕ ***then*** ψ is true in case ψ is true in every one of a set of worlds depending on ϕ. According to the *variably material* theories ((Stalnaker 1968; Stalnaker & Thomason 1970) are examples) ***If*** ϕ ***then*** ψ is true in case ψ is true in *a single* world depending on ϕ. The chief difference between the two is that conditional excluded middle

(6) ***If*** ϕ ***then*** $[\psi \vee \chi] \rightarrow [$***If*** ϕ ***then*** $\psi \vee$ ***If*** ϕ ***then*** $\chi]$

holds in the variably material accounts.

The variably strict theories are much more popular; the variably material theories seem to be much better supported by the linguistic evidence.

Cross' rule represents a variably strict theory of $\boldsymbol{Can}_A$ (or rather, of its negation, $\neg\boldsymbol{Can}_A$). The corresponding variably material theory would make $g(\phi, \mathbf{A}, w)$ a unit set. We can simplify the picture by positing a function f from formulas, agents and worlds to worlds. The satisfaction condition for ability would then be as follows.

(7) $M \models_w \boldsymbol{Can}_A\phi$ if and only if $M \models_{f(\phi,\mathbf{A},w)} \phi$.

The intuitive meaning of f is that $f(\phi, \mathbf{A}, w)$ should be the closest world to w in which $\mathbf{A}$ tries to bring about ϕ.[2] We impose one condition on f.

(8) If $M \models_w \phi$ then $f(\phi, A, w) = w$.

Rule (7) has the effect of validating $\boldsymbol{Can}_A\phi \vee \boldsymbol{Can}_A\neg\phi$; Condition (8) validates $\phi \rightarrow \boldsymbol{Can}_A\phi$. The logic of $\boldsymbol{Can}_A$ can be axiomatized by adding these as axiom schemas to basic axioms for modal operators. One could ask whether the underlying modal logic should be that of **S4** or even **S5**, but I think it is unrewarding to press these details too far. Instead, I will simply note that the reformulated modal theory of ability is closer to situation-based accounts, because f is now analogous to the RESULT function. The analogy will be improved as soon as f is reformulated to apply to actions rather than propositions.

[2]Cross requires that his function g should give the attempt a fair trial—things must be normal with respect to A's attempt to bring ϕ about. I agree that this condition is important for generic ability, but that it is not part of occasional ability. Practical ability has to take adverse circumstances into account.

Providing 'can' with actions as arguments

This project divides into two parts: (i) adopting a logic with a quantificational domain that contains actions (here, I mean individual actions, not action types) and (ii) providing an account of the formalization of commonsense examples involving 'can'.

Two quite different traditions do the first job: eventuality-oriented natural language semantic theories, such as (Parsons 1990; Higginbotham, Pianesi, & Varzi 2000), and action centered dynamic theories such as the Situation Calculus. (Steedman 1998; 1998) combine both traditions. I can rely on the natural language tradition to deal with the connections to natural language issues, and will concern myself from now on with task (i). I will try to focus on reasoning issues in the remainder of this paper, and will mainly draw on the dynamic tradition.

The account that is presented in the following sections will be simplified in many ways. Eventually, I would like to produce a formulation that takes into account full abstraction hierarchies of higher and lower-level actions (that is, hierarchies of actions and hierarchical planning), and aspectual types and actions that are characterized in terms of the results they achieve.

I will use the Situation Calculus as a starting point. This too could be considered to be a simplification; but the Situation Calculus has proved to be remarkably robust. Issues about ability are closely related to knowledge-how, but in this version I will not say much about this aspect of things, which is complicated in its own right. The idea of incorporating knowledge-how into a planning formalism goes back to (Moore 1985).

Situation Calculus

I would like to approach the problem of formalizing ability in a way that could be carried out in most formalisms for reasoning about action and change. But it will be convenient to explain the basic ideas using a version of the Situation Calculus.

I'll use Many-Sorted First-Order Logic as the vehicle of formalization. There is a sort AC of actions, a sort FL of fluents or states, a sort SI of situations, and a sort IN of garden-variety individuals. I'll adopt a convention of flagging the first occurrence of a sorted variable in a formula with its sort. I'll do this with constants also, except that: (1) a is reserved for constants of sort AC, (2) s is reserved for constants of sort SI, (3) f is reserved for constants of sort FL, and (4) c is reserved for constants of type IN.

In the standard SC approach, you have a function $\mathbf{r}$ from actions and situations to situations. $\mathbf{r}(\mathbf{a}, \mathbf{s}) = \mathbf{s}'$ means that $\mathbf{s}'$ is the situation that results from performing $\mathbf{a}$ in $\mathbf{s}$. (The existence of $\mathbf{r}$ presupposes a deterministic sort of change, at least as far as action-induced change goes.) Our formal language contains a function letter RESULT denoting the function $\mathbf{r}$.

If you suppress considerations having to do with causality and the Frame Problem (which I propose to do for the time being) the formalism is pretty simple. Planning knowledge is indexed to actions, in the form of *causal axioms* which associate conventional effects and preconditions with actions.

The causal axiom for an action **a** denoted by a has the following form.

(9) $\forall x_s[\text{PRE}(a, x) \rightarrow \text{POST}(\text{RESULT}(a, x))]$.[3]

Here, PRE is the precondition for a and POST is the postcondition or effect of a.[4]

Situation Calculus with Explicit Ability

Often, a causal axiom of the form Condition (9) is read: "If a is done and $\text{PRE}(a, s)$ is true, then $\text{POST}(a, \text{RESULT}(a, s))$ is true." This provides a perfectly satisfactory basis for reasoning with actions and plans as long as one is only interested in the successful performance of actions. But it is counterintuitive when it may be important to reason about unsuccessful "performances"—i.e., about attempts to perform an action which may fail. This is exactly the sort of reasoning in which "trying" is invoked in informal, commonsense reasoning.

Consider, for instance, a case in which I want to talk to my wife on the telephone. I have a standard method of trying to talk to anyone on the telephone, which consists in (a) finding out the telephone number in case I don't know it, and proceeding to step (b) otherwise; then (b) locating a telephone in case one isn't handy, and proceding to step (c) otherwise; and then (d) dialing the phone number. This method will succeed if my telephone is working and my wife is not using her telephone and is able to answer it; that is, I can call her if these conditions are satisfied. I may know that I can talk to her and proceed to try to talk to her armed with this knowledge; but we seldom act with this sort of assurance. Much more typically, I may not know that I can talk to her but will try to talk to her simply assuming that I can do so, and without any clear fallback plan in case of failure.[5]

To formalize these ideas in a Situation Calculus format, we have to get explicit about trying. This means that we need a more elaborate theory of actions.

Trying

Many commonsense actions[6] are described in terms of the states they achieve; linguists use the term 'telic' for such actions. Generalizing the idea that emerged from the telephoning example, I want to say that the ability to perform such actions consists in (a) having a salient method of achieving the goal of the action in one's repertoire, and (b) being in circumstances such that if the method were acted on, the goal would be achieved.

Although for nontelic actions (jumping, for instance) it is not clear that a method is involved, we do need to invoke a method in some cases; this means that we will have to incorporate some of the ideas of hierarchical planning in the formalism. To simplify things, I'll assume that the hierarchy has a depth bounded by 2. The models of SC that I have in mind, then, will involve a function **Method**. This function inputs an action **a** and a situation **s**, and outputs the agent's "method of choice" for **a** in **s**. This method is a composite action whose components are basic actions; **a** is basic in case for all **s**, $\mathbf{Method}(\mathbf{s}, \mathbf{a}) = \mathbf{a}$. I use the term 'basic' because actions that are fixpoint of **Method** are similar in some ways to the so-called "basic actions" that have been discussed in the literature on the philosophy of action. See, for instance, (Danto 1973).[7]

I am making a number of assumptions here; let me explain them briefly.

(a) The constraint on the values of **Method** formalizes the idea that the depth of the abstraction hierarchy on actions is at most 2.

(b) An agent may have many ways of performing **a** in **s**. I am assuming that one of these is the one that the agent *would* use if it tried to perform **a**. So **Method** has a certain amount of counterfactual content. In cases where planning is involved, the method the agent would use is chosen, and the choice will involve a process of commonsense reasoning. Here, I will only say that I would like to think of this reasoning process as abductive. According to this way of looking at things, the method that is chosen is the series of actions that best explains how the goal will be achieved if the actions are carried out. Instead of minimizing explanation costs in the abduction, however, one wants to minimize the "feasibility" costs that attach to the assumption that the actions can be performed. Feasibility costs involve risk, as well as probability.

(c) An agent may have no way of performing **a** in **s**. In the present draft, I'll deal with this in the simplest possible way, by introducing a null act, NULL, with the understanding that $\mathbf{Method}(\mathbf{a}, \mathbf{s}) = \text{NULL}$ means that **Method** is undefined for **a** in **s**.

(d) We want to avoid having to apply the **Method** function to composite actions; so we assume that **Method** is a homomorphism with respect to composition: $\mathbf{Method}(\mathbf{a;b}, \mathbf{s}) = \mathbf{Method}(\mathbf{a}, \mathbf{s})\mathbf{;Method}(\mathbf{b}, \mathbf{s})$.

Whatever "having a method of choice for performing **a** in **s**" means, it is close to saying that an agent knows how to **a** in **s**; knowing how and ability are intimately connected. I won't explore that connection here.

[3] I write $\text{PRE}(a, s)$ rather than $\text{HOLDS}(\text{PRE}(a), s)$. Uses of HOLDS and VALUE are suppressed in similar ways throughout the rest of this paper.

[4] In planning formalisms, what I am calling preconditions are often separated into two kinds of conditions: the ones that are under the agent's control (which are often called "preconditions") and the ones that are not (which are often called "constraints". I do not distinguish between these two sorts of conditions here.

[5] Typically, people will act on commonsense plans without anything like a proof that they will succeed. All that is seems to be required is the absence of a reason for their failure and relatively low risk or cost of trying to achieve the goal, in relation to the benefits of achieving it.

[6] For instance telephoning someone, crossing a street, opening a door, turning on a light.

[7] The similarity consists mainly in the fact that basic actions, as I have defined them, coincide in many cases with basic actions as Danto defines them. But philosophers of action are primarily interested in the relation between causality and action, and Danto's basic actions are defined in terms of causality. I'm primarily interested in reasoning about how to get things done, so my basic actions are defined in terms of methods.

The **Method** function seems to incorporate an insight from hierarchical planning. As in hierarchical planning, we want to have an abstraction hierarchy over actions, so that actions that appear unitary at a coarser scale can be realized by a series of finer-scale actions. A method of trying to do something has to be a way of doing it that is appropriately specific—that is, it needs to be specific enough to permit a reasonable estimate of whether the action can be performed. I won't attempt to formalize this picture in this paper—that is a large project—but I'll try to make the ideas clearer in terms of an example.

We denote the function **Method** by a function letter METHOD of the planning formalism. With the incorporation into the formalism of a distinction between an action **a** and the action of trying to do **a** in **s**, we can revise the causal axiom of the classical Situation Calculus for a constant a denoting an action as follows: we are interested in the results of *trying* to do an action, rather than the results of doing the action itself.

(10) $\forall x_s[\text{CAN}(a, x) \rightarrow \text{POST}(\text{RESULT}(\text{METHOD}(a, s), x))]$.

Such axioms correspond to platitudes of the sort

(11) If I can open the door, then after I try to open the door the door will be open.

$\text{CAN}(a, s)$ can now be characterized in terms of preconditions and constraints on the action denoted by a. For instance, suppose that a denotes a basic action **a**, and that s denotes **s**. Then the success conditions for **a** in s, i.e. the conditions under which trying to perform **a** (in this case simply performing **a**) will achieve the effects conventionally associated with a are simply the preconditions of **a**.

(12) $\text{CAN}(a, s)$ amounts to $\text{PRE}(a, s)$.

For basic actions, then, we are merely repackaging the usual format for causal axioms.

But things are more interesting if a denotes a nonbasic action. For example, a_0 could denote opening a certain door, and $\text{METHOD}(a_0, s)$ could denote grasping the doorknob, exerting a clockwise turning force on the doorknob, and pushing against the doorknob. Let a_1, a_2 and a_3 denote these three actions. I want the preconditions of pushing against the doorknob in this case to include the door's not being stuck. But these are not preconditions of pushing against the doorknob considered in itself; they only count as preconditions when it is considered as a part of $\mathbf{a}_0$. To deal with this complication, I will generalize PRE to involve three arguments: $\text{PRE}(a, a', s)$ combines the preconditions of **a** in **s** when it is considered as a component of $\mathbf{a}'$. Suppose that $\text{PRE}(a_1, a_0, s)$ is 'The agent is next to the door', that $\text{PRE}(a_2, a_0, s)$ is 'The door isn't locked', and that $\text{PRE}(a_3, a_0, s)$ is 'The door isn't stuck'.[8] Then, in this case, $\text{CAN}(a_0, s)$ amounts to 'The door is neither stuck nor locked'.

This treatment of preconditions complicates the associated reasoning. We can't in general associate a set of preconditions with an action without taking the purpose of the action into account. In the case of pushing against the doorknob, the only precondition would be that the agent is at the door if the purpose is merely to test whether the door is open; but the condition that the door is not stuck would have to be added if the pushing is part of opening the door. I would hope that the general knowledge that is needed to associate these preconditions with component actions can be found in general information about processes (like the door's opening) and standard ways of initiating and maintaining them. I am sure that this is a pervasive and important part of commonsense knowledge and that we somehow we need to formalize it, but at this point I'm not prepared to make specific recommendations about how to do this.

In this simplified treatment there is a straightforward definition of CAN:

(13) $\text{CAN}(a_0^s, s)$ amounts to $\bigwedge\{\text{PRE}(a_i, a_0, s) : a_i \text{ is a component of } \text{METHOD}(a_0, s)\}$.

In the door-opening example, the compiled preconditions of opening the door, with the method described in the example, will be that the agent is at the door, the door is unlocked, and that the door is not stuck. In the example, then, 'I can open the door' amounts to the conjunction of these preconditions, perhaps with the addition of a "hidden" preconditions to cover unforeseen failures. In some cases, we may be able to construct something like a proof of the compiled preconditions; this would amount to something like knowledge that one can perform the action, and it corresponds to the case that classical planning formalisms deal with best. In more general cases, what we want is a reason to think that the plan is worth acting on, rather than a proof, and it is in these more general cases that I believe that abductive methods may be appropriate.

An example sketch

There isn't space here to work up an example with full explicitness. Instead, I will take a case that isn't entirely trivial and explain enough about the formalization to make it plausible that we can go from the commonsense knowledge we typically have to axioms.

Assume that I have come up with the following method for achieving a goal:

1. Go to the bookcase.
2. Find my copy of *Emma*.
3. Take it off the shelf.
4. Go to the chair.
5. Sit down.
6. Turn on the lamp.
7. Read chapter 1 of the book.

The following paragraphs are meant to show that the appropriate knowledge is readily available, and in many cases is available in a form that is fairly general.

(a) Suppose that this scenario takes place in my house. I know the layout of my house. Part of knowing the layout is knowing whether paths are open; assuming it's a normal house, this is just a matter of knowing which passages are closed. Having a method of going from my current location

[8] I'm imagining a door with only one lock, which works by preventing the doorknob from turning.

to the bookcase is a matter of finding a preferred route. This too can be done from a spatial representation of the house. If my preferred route from my location to the bookcase is open, it follows (by default) that I can go to the bookcase.

(b) If I have a recognition criterion for an object[9] and it is in a reasonably small area (like a bookcase)[10] and I am located at that area, it follows (by default) that I can find it.

(c) If I have found a book-sized object on an open shelf and I'm next to it then (by default) I can pick it up.

(d) This is like (a).

(e) If I am at a chair and it is not occupied, then (by default) I can sit in it. Either I remember that this chair is unoccupied (by people or things), or I simply assume that it is unoccupied.

(f) This case is typical of ability-related assumptions about the workings of artifacts. Methods for using artifacts are standardized; these cases can be represented by defaults, or generalizations of defaults that produce minimal costs for certain methods. Here, the relevant default chooses turning on the lamp switch as a method of turning on the lamp. Turning on the switch would itself decompose into a method; for instance, I will have to find the switch. The preconditions of turning on the lamp should compile into something like the following set of conditions: the switch is working, the lamp is plugged in, the bulb is not broken (I assume a lamp with one bulb), the bulb is screwed in completely, and the power is on. Some of these conditions may be verifiable from available sensory information; the rest we usually assume for the sake of getting on with things.

Notice that, even when we have no reason to believe that such conditions actually hold, they may be worth acting on. In the case of turning on the lamp, there are no harsh penalties for failure, and if we bring ability assumptions explicitly into the formalization, we are in a position to explain a failure if it occurs. From this explanation, we may well be able to derive another method that will achieve the goal. So there is a sort of dominance argument for using this method; if I can turn on the light it's the most economical method of doing so, and if I can't the method may help me to decide which alternative method to invoke.

(g) If we don't take into account constraints having to do with reading the chapter within certain time bounds, this is a matter of standing know-how (I know how to read English, *Emma* is written in English), physical ability (I can read 10 point type if I'm wearing my glasses, this copy of *Emma* is in 10 point type), and—another artifact assumption—all the pages of chapter 1 are in this copy of the book. (We all have experienced defectively made copies of books, so this last assumption is a default,)

[9]Recognition criteria are an important part of what we learn when we familiarize ourselves with ordinary objects and types of objects. This is probably why one of the things we typically recall most vividly about a book is the look of its cover.

[10]Typically, we seem to be able to estimate with some confidence how likely we are to succeed in a search for some familiar item. I suspect it may be difficult to formalize the reasoning behind such estimates.

It is easy to see that the appropriate defaults are formalizable. The following circumscription-style axiomatization of the ability requirements of the lamp illustrate the pattern.

$$\forall x \forall y_s [[\mathit{Lamp}(x) \wedge \neg Ab_1(x,y)] \rightarrow \mathit{Working}(\mathit{switch}(x,y),y)]$$
$$\forall x \forall y_s [[\mathit{Lamp}(x) \wedge \neg Ab_2(x,y)] \rightarrow \mathit{Plugged\text{-}in}(x,y)]$$
$$\forall x \forall y_s [[\mathit{Lamp}(x) \wedge \neg Ab_3(x,y)] \rightarrow \neg\mathit{Broken}(\mathit{bulb}(x,y),y)]$$
$$\forall x \forall y_s [[\mathit{Lamp}(x) \wedge \neg Ab_4(x,y)] \rightarrow \mathit{Screwed\text{-}in}(\mathit{bulb}(x,y),y)]$$

This axiomatization style provides a home for a more elaborate theory of abnormalities. Lamps for sale in a store may not be plugged in, so they suffer from the second form of anomaly. If my house is regularly cleaned by someone who unplugs the lamp, then the lamp in the example is regularly anomalous. Expectations about normality may be suspended in some cases for reasons that have more to do with risk reduction than epistemology: this is what happens during a pre-flight check.

Advantages and uses

An abnormality theory axiomatizing the circumstances under which an executed plan may fail can serve to diagnose and explain these failures. To the extent that the abnormality theory is complete, we can expect an abnormality to hold when we experience such frustrations. In special cases, we may be able to design abductive algorithms that produce the set of anomalies that could explain a failure. (See, for instance, (Kautz 1990), (Eiter 2002), (Lin & You 2002).) With additional knowledge, we may even be able to rank the anomalies in order of plausibility.

Such a theory provides the beginnings, anyway, of a formal account of the knowledge that is used in producing excuses. As J.L. Austin showed in (Austin 1956 57), the language of excuses is extraordinarily rich and subtle—this is a strong indication that it corresponds to a well-developed area of commonsense reasoning.

Ethical applications aside, the ability to reason about courses of action that have failed is crucially important for practical reasons. The modification of planning formalisms that is suggested here seems to provide a place for this sort of reasoning in a way that matches to some extent the commonsense language and organization of the relevant knowledge.

I think it is an interesting and important feature of this project that plan abstraction hierarchies occupy a central place in the theory. It is impossible, I believe, to begin to give an adequate account of execution failure without enriching the theory of action with a theory of trying, and this leads at once to a hierarchical formulation. Hierarchical representations are, of course, often used in planning applications, but I believe that they are usually motivated by reasons having to do with modularity and efficiency, of the sort that are usually advanced in connection with abstraction hierarchies. This suggests that plan hierarchies are in principle, at least, dispensable in formalizing reasoning about action, and in fact the ideas of hierarchical planning are not usually incorporated in logical formalizations of planning.

I want to suggest that the more formal work has been able to avoid a hierarchical formulation only because it has used an oversimplified model of action that doesn't apply well to reasoning about the failure of actions. The fact that plan hierarchies provide the missing element is very welcome, since hierarchical planning is widely used, and many complex domains have been formalized. Even if the formalizations are not fully declarative and logical, this work provides reason to hope that the relevant knowledge is there and can be represented.

Although the difference between the two formalisms makes a direct comparison difficult, the account of 'can' that emerged from this simple treatment in the Situation Calculus is similar to the more abstract, conditional account framed in terms of possible worlds semantics. In both cases, to say that an agent can perform an action provides a condition that ensures the successful performance of the action. The advantage of the Situation Calculus treatment is that it provides a representation from which we can recover explicit conditions of success. Nothing, of course, guarantees that these conditions should be anything that the agent can control or even know—in formalizing actions that depend on an element of luck, we may have to resort to unknowable "hidden variables." But in the cases where classical planning algorithms are appropriate, it seems that we can recover useful conditions.

Conclusion

In this draft, I have only tried to formalize the simplest version of an action-and-change formalism capable of explicitly representing ability. These simplifications need to be replaced with something more adequate.

Also, the work of integrating the sort of approach advocated here into logical formalisms for reasoning about action and time remains to be done. I haven't begun to work out things like the relation of ability to things like knowing how, nonmonotonicity in action and change, causality, and eventualities of different aspectual types.

However, I hope that this presentation makes a plausible case for the value of the modifications that are suggested here in accounting for commonsense reasoning about plan execution and plan failure.

References

Austin, J. L. 1956–57. A plea for excuses. *Proceedings of the Aristotelian Society* 57:1–30.

Austin, J. L. 1956. Ifs and cans. *Proceedings of the British Academy* 109–132.

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

Cross, C. B. 1985. *Studies in the Semantics of Modality*. Ph.d. dissertation, Philosophy Department, University of Pittsburgh, Pittsburgh, Pennsylvania.

Cross, C. B. 1986. 'Can' and the logic of ability. *Philosophical Studies* 50:53–64.

Danto, A. C. 1973. *Analytical Philosophy of Action*. Cambridge: Cambridge University Press.

Eiter, T. 2002. On computing all abductive explanations. In Dechter, R.; Kearns, M.; and Sutton, R. S., eds., *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, 62–67. Menlo Park, California: American Association for Artificial Intelligence.

Higginbotham, J.; Pianesi, F.; and Varzi, A. C., eds. 2000. *Speaking of Events*. Oxford: Oxford University Press.

Kautz, H. A. 1990. A circumscriptive theory of plan recognition. In Cohen, P. R.; Morgan, J.; and Pollack, M., eds., *Intentions in Communication*. Cambridge, Massachusetts: MIT Press. 105–133.

Kenny, A. 1976a. Human ability and dynamic modalities. In Manninen, J., and Tuomela, R., eds., *Essays on Explanation and Understanding*. Dordrecht: D. Reidel Publishing Company. 209–232.

Kenny, A. 1976b. *Will, Freedom, and Power*. New York: Barnes and Noble.

Kratzer, A. 1977. What 'must' and 'can' must and can mean. *Linguistics and Philosophy* 1(3):337–356.

Lewis, D. K. 1973. *Counterfactuals*. Cambridge, Massachusetts: Harvard University Press.

Lin, F., and Levesque, H. J. 1998. What robots can do: Robot programs and effective achievability. *Artificial Intelligence* 101(1–2):201–226.

Lin, F., and You, J.-H. 2002. Abduction in logic programming: A new definition and an abductive procedure based on rewriting. *Artificial Intelligence* 140(1–2):175–205.

Meyer, J.; van der Hoek, W.; and van Linder, B. 1999. A logical approach to the dynamics of commitments. *Artificial Intelligence* 113(1–2):1–40.

Moore, R. C. 1985. A formal theory of knowledge and action. In Hobbs, J. R., and Moore, R. C., eds., *Formal Theories of the Commonsense World*. Norwood, New Jersey: Ablex Publishing Corporation. 319–358.

Parsons, T. 1990. *Events in the Semantics of English: a Study in Subatomic Semantics*. Cambridge, Massachusetts: The MIT Press.

Stalnaker, R. C., and Thomason, R. H. 1970. A semantic analysis of conditional logic. *Theoria* 36:23–42.

Stalnaker, R. C. 1968. A theory of conditionals. In Rescher, N., ed., *Studies in Logical Theory*. Oxford: Basil Blackwell Publishers. 98–112.

Steedman, M. 1998. The productions of time. Unpublished manuscript, University of Edinburgh. Available from http://www.cogsci.ed.ac.uk/~steedman/papers.html.

Thomason, R. H. 2003. Working notes on ability. Available at www.umich.edu~rthomaso/documents/action/ability.ps.