Knowledge, Action, and Ability in the Situation Calculus: Extended Abstract

Hector Levesque
Dept. of Computer Science
University of Toronto
Toronto, Ont. M5S 1A4
hector@cs.toronto.edu
and The Canadian Institute for Advanced Research

This is joint work with Yves Lespérance and Fangzhen Lin.

The Cognitive Robotics group at the University of Toronto has been studying various aspects of the reasoning and representation problems involving action and change. The setting is that of a (currently simulated) autonomous robot for which we wish to provide high-level control. One problem that has been a central focus of this work is the *frame problem* [5]. Recently, Reiter, building on the work of Shubert, Haas and Pednault, has proposed a simple solution to this problem in the context of the situation calculus [7]. *In the talk, I will briefly review the situation calculus, Reiter's solution to the frame problem, and the regression method he proposes for reasoning about the effect of actions.* Effort is now underway to generalize this solution along a number of dimensions.

One such generalization concerns *knowledge-producing actions*, such as sensing, whose effect is to change what is known about the current situation. In [8] it is shown how such actions can be accommodated within the situation calculus. The idea is to introduce an accessibility relation over situations into the language, as was done by Moore [6], and treat talk of knowledge as an abbreviation for talk about accessibility. The relationship between knowledge-producing actions and this accessibility relation then is no different from the relation between other more physical actions and the predicates they affect. *In the talk, I will review how this approach to knowledge and knowledge-producing actions inherits Reiter's solution to the frame problem, and also the use of regression for reasoning about knowledge and action.*

One particularly serious limitation of Reiter's solution to the frame problem is that it applies only to deterministic, *primitive* actions. More recent work (see in [4]) has extended the solution to complex actions, including iterative, conditional, and non-deterministic ones. The idea is simply to treat formulas that deal with complex actions as abbreviations for (in some cases, second-order) formulas that deal with primitive actions only. The solution to the frame problem for complex actions then is automatically inherited from the primitive actions. *In the talk, I will briefly review this reduction and show how it leads to a novel programming language for high-level robotic control.*

So here is where we stand: we have a solution to the frame problem that incorporates both knowledge and complex actions provided that is there is no knowledge or complex actions: talk of knowledge must reduce to talk about the underlying accessibility relation, and talk about complex

actions must reduce to talk about its component primitives. In this context, we are currently searching for an appropriate formalization of *ability*, the conditions under which a goal of an agent is achievable.

There are two main motivations for this effort. First, we would like to characterize what an agent would have to believe about another agent if the first one was going to use the second one to help in achieving her goals. This was one of the original motivations for the work of Cohen and Levesque on intention [1, 2]: among the many possible ways of achieving a goal, one is to have another agent do some part of it. One way to characterize this type of delegation involves *commitment and ability*: to bring about some condition, it will be sufficient to get another agent committed to it, and see to it that she is able to achieve it. The Cohen and Levesque work focuses almost exclusively on the first conjunct; we now wish to move to the latter. This also provides a reasonably clear criterion of success: our notion of ability will be adequate if we are able to prove that whenever an agent remains committed to bringing about some condition, and is able to bring it about (according to our definition), then the condition eventually will obtain.

But there is an even more basic reason for wanting a formal account of ability within the situation calculus: we want to state precisely what it means for a robot to be able to achieve a goal. The method we currently use is the classical one due to Green [3]: we pose an existentially quantified description of a goal situation as a theorem to be proved, and we use answer extraction to obtain a sequence of primitive actions which will bring it about. Even though we ensure that the agent is *physically able* to execute each action in the sequence, this account still leads to incorrect predictions about ability. *In the talk, I will show how this classical approach to ability (and even an epistemic version of it) fails to account for what a robot needs to know and when.* The main issue to be addressed is making sure that an agent always knows at each point the next primitive action to perform.

Partly motivated by such concerns, Moore has proposed a definition of ability that appeals to the existence of a complex action (program) [6]. Roughly, an agent is able to achieve a goal if there is a complex action A such that the agent knows that doing A results in a state where the goal is satisfied, and furthermore, the agent *knows how* to do A. Knowing how to do a complex action is in turn defined recursively in terms of the action structure. For example, an agent knows how to execute a conditional if she knows the truth value of the condition and knows how to execute the appropriate branch. In the case of atomic actions, what is required is that the agent know what atomic action is required. *In the talk, I will briefly discuss some advantages and limitations of Moore's account.* One major problem with this definition for us is that it relies on quantifying over complex actions, and as mentioned above, our solution to the frame problem is predicated on having only primitive actions in the domain of discourse.

Another account that captures many of the features of Moore's without quantifying over complex actions is an inductive one. Roughly, an agent is able to achieve ϕ iff there is some n such that the agent is able to achieve ϕ in n steps. We define ϕ as being achievable in 0 steps as knowing that ϕ is already true, and in $n+1$ steps, as knowing a (primitive) action a such that after doing a, ϕ is achievable in n steps. The main drawback to this definition is that it is too *strong*: it forces an agent that is able to achieve ϕ to essentially know how many steps are required. *In the talk, I will present an example where an agent is intuitively able to achieve some condition, without having any idea of how long it will take.* One possible remedy is to weaken the definition and merely require the agent to know at each point that there is a some n (without necessarily knowing what it is) such

that n steps from now ϕ will be true. *In the talk, I will show that this version is too weak.*

A related definition is based on a fixed-point construction. Consider the following two constraints on a predicate C over situations (for a given sentence ϕ):

1. If ϕ is known in situation s, then $C(s)$.

2. If there is a primitive action a such that the agent knows in situation s that C holds of the situation resulting from doing a now, then $C(s)$.

Clearly the property of being able to achieve ϕ in situation s (however it is to be defined) ought to satisfy these two constraints. But many other predicates satisfy them as well, for example, the universal set of situations. However, it is not hard to show that there is a unique *minimal* set that satisfies the two constraints. Thus, we might want to say that an agent is able to achieve ϕ in situation s iff s is an element of this minimal set. Less obvious perhaps than in the iterative case, this definition is once again too strong. *In the talk, I will present an example showing this.*

The definition of ability that we are currently exploring appears to overcome these obstacles, and does so without requiring quantification over complex actions. To define ability, we appeal instead to the notion of a *path* in the situation calculus. *In the talk, I will define the notion of a path and explain why it is independently useful.* First, we define what it means to be able to "follow a path" between two situations; next, we say that an agent can "get to ϕ on a path" if there is a situation on the path such that the agent can follow the path there and, in addition, ϕ holds at that situation; finally, we say that an agent is able to achieve ϕ (or equivalently, can get to a situation where ϕ holds) if there is a path such that the agent knows that she can get to ϕ on that path. *In the talk I will also discuss variants of this definition, including one that appears to be quite useful, where we have the existence of a path without requiring the agent to know about it.*

This definition appears to have a number of interesting properties. For example, it satisfies the above two constraints, while also handling the example where the minimal predicate was too strong. In addition, it can be shown to lie between the overly strong and overly weak iterative accounts. Also, since it does not rely on complex actions, it remains compatible with the solution to the frame problem.

In fact, instead of defining ability in terms of knowing how to do a complex action, we can do the reverse and define knowing how to execute a complex action in terms of ability: we say that an agent knows how to do a complex action A if she is able to get to a final state of A starting now. With this definition, it is possible to prove what was assumed before about Moore's account: if an agent knows how to do A and knows that in any final state of A the condition ϕ holds, then the agent is able to achieve ϕ. *In the rest of the talk, time and space permitting, other properties of this definition will be explored.*

References

[1] P. R. Cohen and H. J. Levesque. Intention is Choice with Commitment. *Artificial Intelligence*, 42(3), 1990.

[2] P. R. Cohen and H. J. Levesque. Teamwork. *Noûs*, 25(4):487–512, 1991.

[3] C. Green. Application of theorem-proving techniques to problem-solving. In *IJCAI-69*, Washington, D. C., May 1969.

[4] Yves Lespérance, Hector J. Levesque, and Fangzhen Lin. A formalization of ability and knowing how that avoids the frame problem. Submitted to KR-94.

[5] J. McCarthy and P. J. Hayes. Some Philosophical Problems from the Standpoint of Artificial Intelligence. In B. Meltzer and D. Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, Edinburgh, Scotland, 1969. Reprinted in Webber, B. L. and Nilsson, N. J., editors, *Readings in Artificial Intelligence*, Tioga Publishing Co., Los Altos, California, pages 431-450, 1981.

[6] R. Moore. Reasoning about knowledge and action. Technical Note 191, Artificial Intelligence Center, SRI International, Menlo Park, CA, 1980.

[7] Raymond Reiter. The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression. In Vladimir Lifschitz, editor, *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy*, pages 359–380. Academic Press, San Diego, CA, 1991.

[8] Richard B. Scherl and Hector J. Levesque. The frame problem and knowledge producing actions. In *AAAI-93*, 1993.