A Note on Logics of Ability

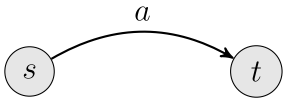
Eric Pacuit and Yoav Shoham

May 8, 2008

This short note will discuss logical frameworks for reasoning about an agent's *ability*. We will sketch details of logics of 'can', 'do', and 'intention'. This note is not meant to be a complete survey of the relevant literature, but rather a starting point for an investigation in logics that can express what actions agents can perform and what state of affairs they can *bring about*. We assume familiarity with basic modal logic, (propositional) dynamic logic and basic temporal logic.

1 Pre-formal Intuitions

The logics that we will explore in this paper all have syntactic objects that are meant to represent an *action*. For example, the propositional dynamic logic (PDL) formula '$[\alpha]\varphi$' is intended to mean "after the action α is performed then φ is true". Already here there are a number of philosophical issues about the nature of *actions* that can influence the formal models we sketch below (see [17] for a survey of the main issues and pointers to relevant literature). Leaving aside much of this philosophical discussion, each of the frameworks that we present below treat actions as *transitions between states, or situations*, as pictured below.



This simple picture describes our basic view of actions: an action relates one state to another state[1].

The primary topic of this note is not only how to formalize that an action a is performed, but also agent-oriented properties. Specifically, we will study logical frameworks that address what it means that an agent *can* do action a, or that

[1] In the above picture, the action a is deterministic (there is only one a-labeled arrow), but this is not necessary for our subsequent discussion — actions may be non-deterministic.

the agent has the *ability to do* action a? A related question is how to formalize that an agent can "bring about" (the truth of) some formula φ, or "see to it that" a formula φ is true. We start with some intuitions about what it means for an agent i to be able to do some action a or bring about some formula φ:

1. **Control**: The action a should be in the control of agent i or the truth of φ should be under i's control.

2. **Repeatability**: Agent i should be able to repeatedly do action a or repeatedly bring about formula φ. Consider a game of darts where a player throws a bulls-eye. Intuitively, we only say that the player has the ability to throw a bulls-eye if he can repeatedly hit the center of the dart board. Alternatively, we can say that the player can *predictably* throw a bulls-eye.

3. **Avoidability**: Agent i should be able to avoid doing action a. So, agents do not have the ability to bring about propositional tautologies.

4. **Free-will**: Agent i should be free to decide whether or not to do action a. That is, the agent should have a choice as to whether or not to do action a or bring about formula φ.

5. **Causality**: Agent i should *cause* action a to take place or the formula φ to become true.

6. **Intentionality**: Agent i should *intentionally* do action a or bring about φ.

Of course, there is a large philosophical literature accompanying each of the last three issues (the relevant entries in the *Stanford Encyclopedia of Philosophy* is a good place to start investigating the literature). The literature on logical frameworks that address the questions raised above is rather difficult to pin down. A special issue of *Stuida Logica* (Volume 51, Numbers 3-4, 1992) edited by Krister Segerbeg contains many relevant article. The first paper in this issue by Segerberg [14] contains an excellent account of a number of relevant formal systems (by now, the discussion is somewhat dated). David Carr has an interesting discussion (in part) about having the physical ability to do an action and *knowing how* to do an action [3]. Finally, the article by Meyer and Veltman [11] from the *Handbook on Modal Logic* contains an account of a number of logical frameworks for reasoning about actions, agency and abilities.

In the following sections we will sketch the details of a few logical frameworks from the literature. Again, our goal is not to provide a comprehensive survey, but rather highlight of few different approaches.

2 General Modal Frameworks

This approach looks for general modal principles that fit our intuitions about actions and agent's abilities. Much of the work here can be viewed as an attempt to formalize various philosophical positions. Here we see familiar modal logics interpreted as logics of action and ability. Consider, for example, the following system proposed by Ingmar Pörn [**?**]. Pörn uses a simple propositional modal language with expressions of the form $Do(A)$ intended to mean "the agent does A", or "brings about A". The logic proposed in the modal logic **T**:

$$\begin{array}{ll} K & Do(A \to B) \to (Do(A) \to Do(B)) \\ T & Do(A) \to A \\ Nec & \text{if } \vdash A \text{ then } \vdash Do(A) \end{array}$$

Formulas are interpreted in the usual Kripke structures. While some of the above axioms are intuitively appealing (eg., the T axioms says that "if the agent does A then A is the case"), there are some less desirable consequences of the above system. In particular all of the following are derivable in the above axiom system (actually, all the following are derivable without the T axiom):

1. If $\vdash A \to B$ then $\vdash Do(A) \to Do(B)$
2. $Do(A) \wedge Do(B) \to Do(A \wedge B)$
3. $Do(\top)$

Note that the last formula is inconsistent with intuition 3 from the previous section: agents should be able to avoid doing an action. Contrary to this principle, formula 3 above says that an agent can always "bring about" a tautology.

Various researchers have investigated variants of the above axiomatization (see [7] for the references). Dag Elgesem identified a core set of axioms [6] in a bi-modal language with CA meaning "the agent is capable of realizing A" and EA meaning "the agent brings about A":

1. All propositional tautologies
2. $\neg C\top$
3. $EA \wedge EB \to E(A \wedge B)$
4. $EA \to A$
5. $EA \to CA$

plus the rules modus ponens and substitution of equivalent formulas (from $A \leftrightarrow B$ we can infer both $EA \leftrightarrow EB$ and $CA \leftrightarrow CB$). The above logic is a *non-normal* modal logic[2] and so the usual Kripke semantics can no longer be used[3]. Governatori and Rotolo [7] investigate the completeness of the above system with respect to various standard approaches of providing semantics for non-normal modal logics including Kripke structures with *impossible worlds*[4] and *neighborhood structures*[5].

3 Dynamic Logic Frameworks

The frameworks here draw on the extensive literature (mostly by computer scientists) on *propositional dynamic logic* [8]. In PDL, (complex) actions are part of the logical language. That is, in standard PDL, formulas and actions expressions are generated by mutual recursion:

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \varphi \mid [\alpha]\varphi$$

$$\alpha := a \mid \alpha ; \alpha \mid \alpha \cup \alpha \mid \alpha^* \mid \varphi ?$$

where p is a propositional variable and a is an atomic action. The formula $[\alpha]\varphi$ means "after executing action α, φ is true". The intended interpretation of the action expressions are as follows:

1. Concatenation: $\alpha ; \beta$. FIrst do α then do β.
2. Nondeterministic Choice: $\alpha \cup \beta$. Either execute α or β.
3. Kleene Star: α^*. Execute α any finite (including 0) number of times.
4. Test: $\varphi ?$. This action does not change the state and succeeds provided φ is true (at the state where it is executed).

[2] A modal logic is normal if it contains the K axiom $\Box(A \to B) \to (\Box A \to \Box B)$ and the rule of necessitation — **K** is the minimal normal modal logic. A modal logic is non-normal if it is not normal.

[3] The K axiom is valid on any Kripke frame and so will be contained in any complete logic over any class of Kripke frames.

[4] An impossible world is a world where all formulas are possible, but no formula is necessary. That is, all formulas of the form $\Diamond A$ are true but all formulas $\Box A$ are false.

[5] Neighborhood models as a semantics for modal logic was invented in the 1970s by Richard Montague and Dana Scott. A neighborhood model generalizes a Kripke structure by assigning to each state w a collection of subsets of states, i.e., the collection of sets that are necessary at w. In a neighborhood model, $\Box\varphi$ is true at w provided the set of states that satisfy φ is in the neighborhood of w (the collection of sets assigned to w). See [12] for complete discussion of neighborhood structures as a semantics for modal logic.

Formulas of PDL are typically interpreted in labeled transition systems where the actions are interpreted as *paths* (see [8] for details). PDL was introduced by Vaughn Pratt to reason about computer programs, but what about interpreting PDL as a logic of *actions* in the sense discussed in this paper?

An early approach to interpret PDL as logic of actions was put forward by Krister Segerberg [13]. Segerberg adds an "agency" program to the PDL language δA where A is a formula. The intended meaning of the program 'δA' is that the agent "brings it about that A'. As usual, the interpretation of δA is a set of paths. Segerberg interprets the action δA as the set of all paths p in a fixed labeled transition systems such that

1. p is the computation according to some program α, and

2. α only terminates at states in which it is true that A

Interestingly, Segerberg also briefly considers a third condition:

3. p is optimal (in some sense: shortest, maximally efficient, most convenient, etc.) in the set of computations satisfying conditions (1) and (2).

But this last condition is not pursued. Segerberg extends the usual PDL axiomatization[6] with the following two axioms

1. $[\delta A]A$

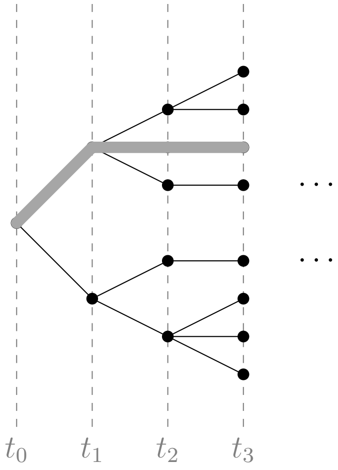
2. $[\delta A]B \to ([\delta B]C \to [\delta A]C)$

This general framework has been used as a logic for reasoning about agents and their abilities (see, for example, [4] and [5] for discussions).

4 Temporal Logic Frameworks

Alternative accounts of agency do not include explicit description of the actions as in the previous section. Instead the frameworks are based on a branching-time structure. That is, formulas are interpreted in trees where branching is interpreted as a choice for the agent. Consider the picture below:

[6] Actually, the usual axioms of PDL was proposed by Segerberg and proved complete with respect to the standard semantics by Rohit Parikh.



Each node represents a choice point for the agent. A **history** is a maximal branch in the above tree. Formulas are interpreted at history moment pairs. For example, let h be the highlighted history above, then we may interpret formulas at (h, t_1). The key modality is $[stit]\varphi$ which is intended to mean that the agent i can "see to it that φ is true". There are different definitions proposed for interpreting the above modality. One example is *deliberative stit*, where $[stit]\varphi$ is true at a history moment pair provided the agent can choose a (set of) branch(es) such that every future history-moment pair satisfies φ *and* there is a choice agent i can make in which φ becomes false[7]. This framework is explored (among other places) in the book by Belnap, Perloff and Xu [2] and more recently by Horty in [9]. Complete axiomatizations can be found in the article by Xu [18] and more recently by Balbiani, Herzig, and Troquard [1].

We end this section with a brief discussion (based on [9, pg. 20-21]) to illustrate some of the main ideas of this framework. We first need some notation. If T is a tree (as above) and t is a moment, then H_t is the set of histories going through t. A choice at moment t, denoted $Choice_t$ is a partition of H_t with $Choice_t(h)$ the partition cell in $Choice_t$ containing h. Then $[stit]\varphi$ is true at history moment h/t provided $Choice_t(h) \subseteq |\varphi|_t$ and $|\varphi_t| \neq H_t$. Formulas are interpreted at history/moment pairs and we write $|\varphi|_t = \{h \in H_t \mid \varphi \text{ is true at } h, t\}$. We use the modality '$\Diamond$' to mean historic possibility. That is $\Diamond\varphi$ is true at a history moment pair h, t provided there is some history $h' \in H_t$ such that φ is true at h', t. Horty considers the following definition individual ability: an agent has the ability to bring about φ at a history moment h/t provided $\Diamond[stit]\varphi$ is true at h/t.

Consider the example of an agent (call her Ann) throwing a dart. Suppose Ann is not a very good dart player, but she just happens to throw a bull's eye.

[7] There is another proposal in the literature, called *cstit* (which stands for Chellas-*stit*, named after Brian Chellas) which is true provided the agent can make a choice to guarantee φ is true.

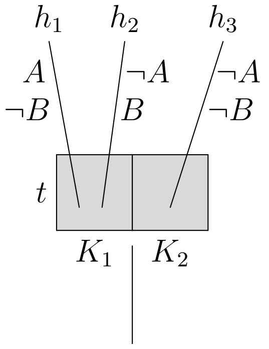
Intuitively, we do not want to say that Ann has the *ability* to throw a bull's eye even though it happens to be true. That is, the following principle should be falsifiable:

$$\varphi \to \Diamond[stit]\varphi$$

Continuing with this example, suppose that Ann has the ability to hit the dart board, but has no other control over the placement of the dart. Now, when she throws the dart, as a matter of fact, it will either hit the top half of the board or the bottom half of the board. Since, Ann has the ability to hit the dart board, she has the ability to either hit the top half of the board or the bottom half of the board. However, intuitively, it seems true that Ann does not have the ability to hit the top half of the dart board, and also she does not have the ability to hit the bottom half of the dart board. Thus, the following principle should be falsifiable:

$$\Diamond[stit](\varphi \vee \psi) \to \Diamond[stit]\varphi \vee \Diamond[stit]\psi$$

The following model will falsify both of the above formulas:



Here, for example, A is true at h_1, t and false at h_2/t and h_3/t, so $|A|_t = \{h_1\}$. The agent's choice at t is $Choice_t = \{K_1, K_2\} = \{\{h_1, h_2\}, \{h_3\}\}$. We leave it to the reader to check that both $A \to \Diamond[stit]A$ and $\Diamond[stit](A \vee B) \to \Diamond[stit]A \vee \Diamond[stit]B$ are false at h_1/t.

5 The Situation Calculus

The situation calculus is a framework for talking about knowledge and ability of agents popular in the AI literature (it was introduced by John McCarthy in 1969). Formally, the situation calculus is a second-order many-sorted logic (see, for example, [10] for a discussion of the framework and [16] for a comparison between the situation calculus and related modal frameworks). A typical formula in the situation calculus is $Result(s, a) = s'$ which says that "the[8] result of executing action a in situation s is situation s'".

For this paper, our interest is with the expression $Can(s, \varphi)$ which is intended to mean that "the agent can achieve φ in situation s". A possible definition is:

$$Can(s, \varphi) \;:=\; \exists a(Result(s, a) \text{ satisfies } \varphi)$$

Here 'a' may be a complex action. So, $Can(s, \varphi)$ holds if in situation s there is some (possibly complex) action that results in a state satisfying φ. Hector Levesque and colleagues offer the following refinement of this definition of what an agent *can* do in a situation [10]. The main idea is that it is not enough for there to exist an action that leads to a situation satisfying φ, in addition the agent must *know* that the action leads to a state satisfying φ. A fixed-point definition is offered to make this intuition formal: $Can(s, \varphi)$ is true provided

1. if $K(s, \varphi)$ (the agent knows φ in situation s) is true
2. if $\exists a K(Result(s, a), Can(Result(s, a), \varphi))$

This definition raises an interesting question: *what does it meanfor an agent to know how a plan will achieve its goal*[9]?

6 Logics of Intention

There are a variety of logical frameworks for reasoning about not only agent's knowledge/beliefs and abilities but also *motivational attitudes* such as *intentions*. We will not discuss these frameworks here and refer to the textbook [15, Section 14.4] for details.

References

[1] BALBIANI, P., HERZIG, A., AND TROQUARD, N. Alternative axiomatics and complexity of deliberative STIT theories. *Journal of Philosophical Logic* (2007).

[2] BELNAP, N., PERLOFF, M., AND XU, M. *Facing the Future.* Oxford University Press, 2001.

[3] CARR, D. The logic fo knowing how and ability. *Mind 88* (1979), 394 – 409.

[8] Again, we are assuming actions are deterministic but this is not necessary.

[9] And, of course, how do we formalize it?

[4] Chellas, B. On bringing it about. *Journal of Philosophical Logic 24* (1995), 563–571.

[5] Eglesem, D. Intentions, actions and routines: A problem in krister segerberg's theory of action. *Synthese 85*, 1 (1990), 153–177.

[6] Elgesem, D. The modal logic of agency. *Nordic Journal of Philosophical Logic 2*, 2 (1997), 1 – 46.

[7] Governatori, G., and Rotolo, A. On the axiomatization of elgesem's logic of agency and ability. *Journal of Philosophical Logic 34*, 4 (2005), 403–431.

[8] Harel, D., Kozen, D., and Tiuryn, J. *Dynamic Logic.* The MIT Press, 2000.

[9] Horty, J. *Agency and Deontic Logic.* Oxford University Press, 2001.

[10] Lesperance, Y., Levesque, H., Lin, F., and Scherl, R. Ability and knowing how in the situation calculus. *Studia Logica 66* (2000).

[11] Meyer, J.-J., and Veltman, F. *Handbook of Modal Logic.* Elsevier, 2007, ch. Intelligent Agents and Common Sense Reasoining.

[12] Pacuit, E. ESSLLI course on neighborhood semantics for modal logic. Course notes found at `ai.stanford.edu/∼epacuit/nbhd_esslli.html`, 2006.

[13] Segerberg, K. Bringing it about. *Journal of Philosophical Logic 18* (1989), 327–347.

[14] Segerberg, K. Getting started: Beginnings in the logic of action. *Studia Logica 51*, 3-4 (1992), 347 – 378.

[15] Shoham, Y., and Leyton-Brown, K. *Multiagent Systems.* Cambridge University Press, 2008.

[16] van Benthem, J. Situation calculus meets modal logic. Forthcomin in a collection dedicated to John McCarthy. Available at `staff.science.uva.nl/∼johan`, 2007.

[17] Wilson, G. Action. In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Spring 2008.

[18] Xu, M. On the basic logic of STIT with a single agent. *Journal of Symbolic Logic 60*, 2 (1995), 459–483.