Incorporating Belief Change in a Theory of Action

James Delgrande
School of Computing Science
Simon Fraser University
Burnaby, BC V5A1S6, Canada
jim@cs.sfu.ca

Introduction and Motivation

The overall goal of this work is to try to bring together work in belief change, as represented by the AGM approach (Alchourrón, Gärdenfors, and Makinson 1985), and work in reasoning about action, specifically the situation calculus (Reiter 2001).

On the one hand, the AGM approach has been around for a while, and provides the dominant framework for belief change; however there has been little done in successfully applying this approach in realistic situations or with respect to applications. By locating the AGM approach in the situation calculus one can take into account the fact that the world is dynamic, that it evolves as the result of physical and sensing actions, and the like.

On the other hand, the (basic approach to the) situation calculus assumes a great deal about the domain: that actions always work as intended, that unforeseen events don't arise, that sensors work, etc. Clearly, actions may sometimes fail or have unexpected outcomes, the agent must deal with possibly incorrect information, sensors may not work as advertised, etc. So the goal here is to try to come up with a qualitative theory of an agent in a dynamic world where things may not go as expected, where the agent may learn new information by being told something or by sensing, and where the agent must maintain its stock of beliefs as best it can.

In the case of the situation calculus there has been substantial work on generalising the original formalism. The proposed work builds on (Delgrande and Levesque 2012; 2018); this work in turn derives from (Shapiro et al. 2000) (which in turn generalises (Scherl and Levesque 2003)). Other related work includes, notably, (Fang and Liu 2013; Schwering, Lakemeyer, and Pagnucco 2017).

Components of the Approach

The Situation Calculus The situation calculus is (essentially) a first-order theory in a sorted language with equality. A *situation* is a finite world history, described by a sequence of actions from some initial state of the world. An action takes a state of the world to another state of the world in which the action effects hold. For example, the *fluent instance* $Holding(r, o, s)$ could be used to assert that robot r is $Holding$ object o in situation s, and $\neg Holding(r, o, do(putdown, s))$ would assert that o isn't held following a $putdown$ action.

Scherl and Levesque (Scherl and Levesque 2003) axiomatise an agent's knowledge by treating situations as possible worlds. Two distinguished fluents are used, *SF* and B. The former (mnemonically *sense fluent*) is used for sensing, while the latter provides an accessibility relation between situations. The B fluent is the usual belief accessibility relation: $B(s', s)$ holds when the agent in situation s thinks that situation s' might be the actual situation. Belief is defined, as usual, as truth in all accessible situations.

Belief Revision The standard semantic model (Katsuno and Mendelzon 1991) for AGM-style belief revision takes an agent's epistemic state as being modelled by a total preorder over possible worlds. The least worlds in the preorder characterise the agent's beliefs, $\mathcal{K}$. Then the revision of $\mathcal{K}$ by a formula ϕ, $\mathcal{K} * \phi$ is characterised by the least ϕ worlds in the total preorder. What the total preorder resulting from $\mathcal{K} * \phi$ should look like is a subtle and generally-unresolved question, but suffice to say that various schemes have been proposed, of which we have used those of (Darwiche and Pearl 1997; Nayak 1994). In place of a total preorder, we use *plausibility rankings* (Spohn 1988), in which worlds are assigned a non-negative integer, such that some world has ranking 0.

Belief Revision in the Situation Calculus In previous work, we combined the above two approaches by working with a plausibility ranking over situations. Situations with rank zero characterise the agent's beliefs; those of non-zero characterise propositions the agent believes are false, ranked by their plausibility. In the approach, an agent intending to execute one action might inadvertently execute another. Thus in pushing a light switch, an agent might expect to push the correct switch but could inadvertently push a neighbouring switch. These possibilities were kept track of in the plausibility ordering, so that if an agent later determined that it couldn't have pushed the correct switch (say, by sensing) then it could revise its beliefs to determine the most plausible situations. As another example, see Figure 1, which involves flipping a coin. In the initial situation S_0 on the left hand side, the coin shows heads (H). The agent execute a

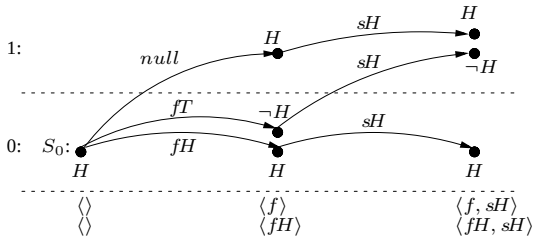

Figure 1: Flipping a coin

flip (f) action, which could be a flip-heads (fH) or a flip-tails (fT), but in reality is a flip-heads. This is illustrated in the middle column: the agent doesn't know which it executed, and it also tracks the possibility that (implausibly) the flip action failed ($null$). The agent then senses (sH) that the coin is heads and so comes to believe that it shows heads. Thus a major concern is keeping track of what an agent believes it executed from what it in fact executed. The two primary predicates in the theory are as follows:

- $B(s', p, s'', s)$ is a fluent that asserts that according to situation s, where the agent believes its actions took it to s'', situation s' has plausibility p. This generalises the standard notion of accessibility.
- $Alt(a_1, a_2, p, s)$ asserts that in situation s an agent intending to execute action a_1 may with plausibility p execute a_2. For example the Alt axiom for flipping a coin can be given by:

$$Alt(f, a, p, s) \equiv$$
$$((a = fH \vee a = fT) \wedge p = 0) \vee (a = null \wedge p = 1)$$

Modifying the theory There are a number of ways in which we plan to modify or enhance the theory.

- The 4-place B fluent leads to a rather intricate successor state axiom. What would be preferable would be to express this as a 3-place fluent, where $B(s', p, s)$ asserts simply that according to situation s, situation s' has plausibility p. This may be doable by modestly restricting things so that an action has two (disjoint) sets of parameters: those controllable by the agent and those by "nature". Then for example flipping a coin could be expressed by an action $flip(x)$, where an agent can execute a $flip$, but here the parameter x is controlled by nature: either heads or tails, each with plausibility 0, or failure with plausibility 1.
- In the approach to date, a sensing action either succeeds with plausibility 0, or fails with plausibility 1. This will be extended to a full theory of sensor failure, using Alt.
- The theory to this point is purely representational. Computational issues, such as dealing with projection, or embedding the approach in an extension of (presumably) GOLOG, are clearly desirable.
- Other possibilities include working in a multiagent environement or also introducing quantitative notions of uncertainty.

Potential contributions

As mentioned, the approach is foremost representational; as such it arguably provides a declarative account of what may "go wrong" in reasoning in a dynamic world: actions may not do what's expected, the agent's beliefs may be incorrect, sensing may be fallible, and so on. In earlier work, the approach has been suggested to provide an epistemic account of *nondeterminism*. That is, the world is assumed to be deterministic, and so in flipping a coin, an agent has either executed a flip-heads or flip-tails; however, the agent doesn't know which it executed, and so believes only that the coin shows either heads or tails. Similarly, the approach may provide a suitable resolution to the *qualification problem*, axiomatising via Alt how things may go amiss, or how actions may simply fail for no known reason. The approach may also prove suitable as a basis for *diagnosis*. For example, consider an agent that believes that a light is on, that toggles the light switch twice, and observes the light is off. Depending on attached plausibilities, it may decide that its initial beliefs were wrong, or that one of the two toggle actions failed.

References

Alchourrón, C.; Gärdenfors, P.; and Makinson, D. 1985. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic* 50(2):510–530.

Darwiche, A., and Pearl, J. 1997. On the logic of iterated belief revision. *Artificial Intelligence* 89:1–29.

Delgrande, J. P., and Levesque, H. J. 2012. Belief revision with sensing and fallible actions. In *Proc. KR 2012*.

Delgrande, J. P., and Levesque, H. J. 2018. An epistemic approach to nondeterminism: Believing in the simplest course of events. *Studia Logica*. To appear.

Fang, L., and Liu, Y. 2013. Multiagent knowledge and belief change in the situation calculus. In *Proc. AAAI*, 304312.

Katsuno, H., and Mendelzon, A. 1991. Propositional knowledge base revision and minimal change. *Artificial Intelligence* 52(3):263–294.

Nayak, A. 1994. Iterated belief change based on epistemic entrenchment. *Erkenntnis* 41:353–390.

Reiter, R. 2001. *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. Cambridge, MA: The MIT Press.

Scherl, R., and Levesque, H. 2003. Knowledge, action, and the frame problem. *Artificial Intelligence* 144(1-2):1–39.

Schwering, C.; Lakemeyer, G.; and Pagnucco, M. 2017. Belief revision and projection in the epistemic situation calculus. *Artificial Intelligence* 251:62–97.

Shapiro, S.; Pagnucco, M.; Lesperance, Y.; and Levesque, H. J. 2000. Iterated belief change in the situation calculus. *Proc. KR 2000*, 527–538.

Spohn, W. 1988. Ordinal conditional functions: A dynamic theory of epistemic states. In Harper, W., and Skyrms, B., eds., *Causation in Decision, Belief Change, and Statistics*, volume II. Kluwer Academic Publishers. 105–134.