Actions on belief

Sam Steel
Dept Computer Science, University of Essex
Colchester CO4 3SQ, UK
sam@essex.ac.uk

4 Jan 93

Abstract

This paper shows how to represent actions on belief (such as deducing and evaluating) in a language in which it is also possible to represent actions on the world (such as putting one block on top of another). It is done by combining modal logics of belief and of action in such a way that actions on belief can be represented as perfectly respectable modals with standard and well-motivated semantics, obeying sensible rules.

One basic actions is proposed, representing finding that a fact S is true; it has a variant, representing finding that S is true given a fact (or theory) T. These can be used to define other, more usable, actions: finding whether S, finding an object with property ϕ, and evaluating a term.

1 Introduction

1.1 Motivation

This paper is about representing actions on belief, such as deducing and evaluating, in a language in which it is also possible to represent actions on the world such as putting one block on top of another. Why is this interesting?

The first reason is that representing actions on belief is still not properly understood. The second is that I have a long-term programme of building a reflexive planner — that is, a planner that is able to plan to plan as well as plan to act. Such a planner would need to be able to reason about actions on belief — deciding what to do — as well as actions on the world. I want to show that such actions are legitimate and well-behaved. I shall try to do that by showing how modal logics of belief and of action can be combined. I can then describe what it is for an action to be informative or to lose information. The basic idea turns out to be that of Moore (Moore 85), but re-expressed and extended. An action preserves information or ignorance iff it obeys certain simple rules.

So one can describe the effect of a given action on the agent's information. Is it also possible to characterize actions purely by their effect on the agent's information? For instance, can one describe the action of finding whether S is true or not, independently of how one does so? The answer is "yes". The central idea is that the set of states of affairs compatible with what an agent believes may be split or enlarged by an action. Some such splittings lead to the agent ceasing to be in doubt about S. One can combine all splitting in such a way that the combination obeys the rules that one would expect of an action of discovery. It then becomes possible to describe the actions of finding an object with a certain property, and of evaluating a complex term, in the same way.

The relation of this work to Moore's is:

- The metaphysics of state, belief and change is the same. Moore presents this by working in the meta language of epistemic and dynamic logic; this paper works more at the object level. This difference is minor.
- It offers an account of the preservation across action of information and ignorance of which Moore's account is a sub-part.
- It introduces semantically respectable actions that gain belief by constructing them as a "limit" of the set of all information-preserving actions.

One standard objection to logicist accounts of mental actions is logical omniscience. If an answer to a question can be found by inference from a given theory, and if our mental actions are logical, then surely we must know the consequences of the theories we accept; but we don't know them; so (it is claimed) our mental processes cannot be logical.

There is an alternative; if agents are taken not to believe theories but instead to try to answer questions in the light of theories (in a sense that can be clearly defined) then action on belief can be persuasively described without accepting logical omniscience. I stress that this paper make no claims about the psychological processes of real agents; all it says is of the form "Agents can be described as if ...".

This paper is an essay in applied philosophical logic. I had some intuitions about the metaphysics of belief and action, and tried to capture them. I have tried throughout to capture them in the semantics, to find properties of models that look right and are simple to describe, and only then have I looked to see what sentences they validate.

Descriptions of logical systems are expected to include sets of axioms and a proof of their completeness. This paper does not. My interest has been in persuasive analysis of the problem, and in rules that can be put into an automatic reasoning system. That has led to an interest in valid sentences rather than in axioms. Since deductive completeness is not the effective limitation on the power of automated reasoning, I have not pursued completeness either.

The paper reviews epistemic and dynamic logic, and introduces some minor extensions to them. It then classifies the ways action and belief can interact,

and considers the use of litmus. Then the first action characterized entirely by its effect on belief, **finding that**, is described, and actions of **finding whether**, **finding a** and **evaluat**ing are built out of it. Then similar actions done in the light of a theory but without logical omniscience are described. Lastly there is a sketch of how this applies to reflexive planning.

1.2 Revision of epistemic and dynamic logic

Representing action on belief involves representing belief and action. I shall do this in modal logic, since there is already a modal logic of belief (epistemic logic) and of action (dynamic logic). It is easy to combine them since their syntax and semantics are so similar.

I assume the reader is familiar with epistemic and dynamic logic (Pratt 76), (Hintikka 1962), (Goldblatt 1987). (There is a review of the syntax and a standard Kripke semantics below.) I shall however recall the application of modal logic to belief and action.

The central idea of epistemic logic is this: suppose I believe that there is a block on a table, but that I do not know whether it is red or green. That uncertainty about the world is to be represented, not by describing a state of affairs which says nothing about the block's colour, but by describing two complete states of affairs, where in each there is a block on the table, but where in one the block is red, and in the other, green. Only what is true in every state of affairs described is believed. Gaining certainty is then cutting down on uncertainty; some of the possible states of affairs are ruled out, so the number of facts true in every state of affairs goes up.

In epistemic logic, that is formalized by representing the actual state of affairs by an index — a point associated with a first-order model of the world — and by representing the states of affairs that the agent finds credible at any index by other indices linked to it by an epistemic accessibility relation, *BEL*. (**Bel** is a modal, *BEL* is the relation it denotes.) Those indices can then be linked to others, and so on. The "information set" at any index i is defined as the set

$$\{ \text{j} \mid \langle \text{ij} \rangle \in BEL \}$$

A sentence S is believed at index i iff S is true at every index in the information set. One knows what a term T means iff T denotes the same object at every index in the information set.

The central idea of dynamic logic is to keep the idea of indices representing states of affairs, but to let relations on them represent actions. Suppose the actual state of affairs corresponds to some index, and that action A is done. The indices reached by the relation denoted by A correspond to the states that are possible outcomes of doing A. There may be zero, one or many such indices. A sentence S is an effect of an action A at index i iff S is true at every index in the image of i under (the relation denoted by) action A.

Given a point i and a relation R, I shall call the set

$$\{ j \mid \langle ij \rangle \in R \}$$

the "image" of i under R, by analogy with the image of a point under a function. The image of a set X under R is

$$\{ j \mid \exists i.\ i \in X \ \&\ \langle ij \rangle \in R \}$$

If T is a term, it makes sense to talk about "the thing that is T after doing A" iff T denotes the same object at every index accessible by the relation denoted by A. It may be the case that if a sentence P is true before an action A, then a sentence Q is true after it. That can be written

$$P \rightarrow [A]\ Q$$

and is the way of expressing the fact that P is a precondition of Q being true after A.

An action that doesn't take place is taken to denote the empty relation on indices. In particular, **abort** denotes the empty relation. An action may not take place because it is impossible in the circumstances in which it is attempted, or because it fails to terminate. No states are reached by an impossible action, so, trivially, anything is true after it, even falsity. Indeed, impossible actions can be characterized by the fact that they satisfy

$$[A]\ \#$$

1.3 The object language

The conceptual content of this section is slight. It reviews and combines the languages of epistemic and dynamic logic, and introduces some extensions to the object language of modal logic needed later. The first part is completely standard; the extensions are described later.

The language is first-order. Missing connectives come via the usual definitions. The extension to non-unary predication and application is trivial. The $*$ operator of dynamic logic is not needed.

There is a small snag because **Bel** S and [**A**] S mean the same sort of thing, but only the action modal needs the [...]. Allow **Bel** S as an abbreviation.

Sentence ::= Predicate(Term)
| $\text{Sentence}_1 \rightarrow \text{Sentence}_2$
| #
| ∀Var. Sentence
| [Modal] Sentence
| **Bel** Sentence
Term ::= Constant | Var | Function(Term)
Modal ::= **Bel**

| Action
| Sentence?
| Modal_1 ; Modal_2
| $\text{Modal}_1 \cup \text{Modal}_2$

A model M is a structure

$\langle\langle \text{Indices}, \text{Dom}, \mathit{BEL}\rangle, \text{VC}, \text{VA}\rangle$

There is a set of indices (states, worlds) Indices. Dom is the domain of objects, for convenience assumed to be the same at all indices. *BEL* is the epistemic accessibility relation on Indices. VC interprets terminal symbols of the categories Constant, Predicate and Function as members of, subsets of and functions in the domain in the usual way; but the interpretation can vary from index to index. VA interprets primitive Actions as relations on Indices.

$\models$ S iff for all models M. for all assignment functions g. M g $\models$ S

M g $\models$ S iff for all i $\in$ Indices of M. M g i $\models$ S

M g i $\models$ Predicate(Term) iff $[\![$ Term $]\!]$ M g $\in$ VC(i)(Predicate)

M g i $\models$ $S_1 \rightarrow S_2$ iff M g i $\models$ S_1 implies M g i $\models$ S_2

M g i $\models$ # is false

M g i $\models \forall$Var. S iff for all d $\in$ Dom. M g[Var:=d] i $\models$ S

M g i $\models$ [Modal] S iff for all j ($\langle$ij$\rangle \in$ $[\![$ Modal $]\!]$ M g $\rightarrow$ M g j $\models$ S)

M g i $\models$ **Bel** S iff M g i $\models$ [**Bel**] S

For the terms

$[\![$ Constant $]\!]$ M g i $\triangleq$ VC(i)(Constant)

$[\![$ Var $]\!]$ M g i $\triangleq$ g(Var)

$[\![$ Function(Term) $]\!]$ M g i $\triangleq$ ($[\![$ Function $]\!]$ M g i)($[\![$ Term $]\!]$ M g i)

For the modals (all of which are independent of the index at which they are evaluated)

$[\![$ **Bel** $]\!]$ M g $\triangleq$ *BEL*

$[\![$ Primitive $]\!]$ M g $\triangleq$ VA(Primitive)

$[\![$ Sentence? $]\!]$ M g $\triangleq$ { $\langle$ii$\rangle$ | M g i $\models$ Sentence }

$[\![$ A_1 ; A_2 $]\!]$ M g $\triangleq$ $[\![$ A_1 $]\!]$ M g ; $[\![$ A_2 $]\!]$ M g

where ; is "diagram-order" composition

R ; S $\triangleq$ { $\langle$xz$\rangle$ | $\exists$y. $\langle$xy$\rangle \in$ R & $\langle$yz$\rangle \in$ S }

$$[\![A_1 \cup A_2]\!] M\ g \triangleq [\![A_1]\!] M\ g \cup [\![A_2]\!] M\ g$$

Some constant actions are

$$[\![\textbf{abort}]\!] M\ g \triangleq [\![\#?]\!] M\ g = \{\}$$

$$[\![\textbf{skip}]\!] M\ g \triangleq [\![(\# \rightarrow \#)?]\!] M\ g = \{\ \langle ii \rangle \mid i \in \text{Indices}\ \}$$

This semantics supports the axioms found in any "normal" modal logic

$$\text{if} \models S \text{ then} \models \text{Modal}\ S$$

$$\models \text{Modal}\ (S \rightarrow T) \rightarrow (\text{Modal}\ S \rightarrow \text{Modal}\ T)$$

and the standard dynamic logic axioms.

$$\models [A]\ [B]\ S \equiv [A\ ;\ B]\ S$$

$$\models [A]\ S\ \&\ [B]\ S \equiv [A \cup B]\ S$$

$$\models [S?]\ T \equiv (S \rightarrow T)$$

Since modals are relations, and relations are sets of pairs of indices, one can say that one modal is included in another, in the sense that all the arcs leaving the index in one modal are also arcs in the other modal. To express that in the object language, add

$$\begin{aligned} \text{Sentence} ::= &\ \text{Modal}_1 \stackrel{\bullet}{\supseteq} \text{Modal}_2 \\ \mid &\ \text{Modal}_1 \subseteq \text{Modal}_2 \\ \mid &\ \text{Modal}_1 \stackrel{\bullet}{=} \text{Modal}_2 \end{aligned}$$

The semantics of the first is given by

$$\begin{aligned} M\ g\ i \models \text{Modal}_1 \stackrel{\bullet}{\supseteq} \text{Modal}_2 \text{ iff} \\ \forall j.\ (\langle ij \rangle \in [\![\text{Modal}_1]\!] M\ g \leftarrow \langle ij \rangle \in [\![\text{Modal}_2]\!] M\ g) \end{aligned}$$

It is easy to show that that validates

$$\models \text{Modal}_1 \stackrel{\bullet}{\supseteq} \text{Modal}_2 \rightarrow [\text{Modal}_1]\ S \rightarrow [\text{Modal}_2]\ S$$

The clauses and rules about reverse inclusion and equality are analogous.

It is standard that the union of two modals is governed by

$$\text{Modal} ::= \text{Modal}_1 \cup \text{Modal}_2$$

$$[\![\text{Modal}_1 \cup \text{Modal}_2]\!] M\ g \triangleq [\![\text{Modal}_1]\!] M\ g \cup [\![\text{Modal}_2]\!] M\ g$$

That validates

$$\models [\text{Modal}_1]\ S\ \&\ [\text{Modal}_2]\ S \equiv [\text{Modal}_1 \cup \text{Modal}_2]\ S$$

The union of modals can be generalized to the first-order case. This is novel, though slight.

Modal ::= ∪x. Modal

$$[\![\cup x.\ \text{Modal}]\!]\ M\ g \triangleq \bigcup d \in \text{Dom}.\ \ [\![\text{Modal}]\!]\ M\ g[x:=d]$$

That validates

$$\models \forall x.([A]\ S) \equiv [\cup x.\ A]\ S \text{ if } x \text{ is not free in } S$$

Proof: see appendix A.1

The modal "invariably" is to express the idea of a fact being true in every possible world, credible or not.

Sentence ::= **invariably** Sentence

$$[\![\textbf{invariably}]\!]\ M\ g \triangleq \text{Indices} \times \text{Indices}$$

Then

$$\models \textbf{invariably} \dot{\supseteq} \text{Modal}$$

$$\models \textbf{invariably} \dot{\supseteq} \textbf{Bel}$$

and hence

$$\models (\textbf{invariably}\ S) \rightarrow [\text{Modal}]\ S$$

$$\models (\textbf{invariably}\ S) \rightarrow \textbf{Bel}\ S$$

and (by taking Modal to be **skip**)

$$\models (\textbf{invariably}\ S) \rightarrow S$$

"**invariably**" is a means of expressing facts that are not logical truths, but which it is inconceivable should be false; perhaps they are some sort of physical necessity, or the consequences of some indubitable theory.

It is useful to have lambda abstracts, merely to express complex predicates. They will not be quantified over or beta-reduced, so they are just a syntactic convenience.

Predicate ::= λVar. Sentence

$$[\![\lambda \text{Var. Sentence}]\!]\ M\ i\ g \triangleq \lambda d \in \text{Dom}.\ \ [\![\text{Sentence}]\!]\ M\ i\ g[\text{Var}:=d]$$

Beta-reduction is not sound if it involves constants outside inside modals replacing variables inside them. For instance, suppose Mary loves the owner of Fido, who is Tom. She still loves Tom after he gives Fido to Fred. So it is true that

(λx. [(tom)gives(fido)to(fred)] loves(mary,x)) (owner(fido))

But that does not entail the naive beta-reduction

[(tom)gives(fido)to(fred)] loves(mary,(owner(fido)))

which means that Mary loves Fred. This can be fixed (Steel 90) but it need not be; potential beta-reductions that would run into that problem can be left unreduced. The only cost will be longer sentences.

2 Interaction of action and belief

This section describes how any action can affect belief.

An epistemic accessibility relation and a relation denoted by an action are both relations on a single set of indices. It is then possible to say that after an action an agent will or will not believe some sentence. In the following diagram, for instance, it is true at index i that after doing A, the agent will believe that S, even though he did not believe it to start with.

Figure 1

That verifies

i $\models \neg$(**Bel** S) & $\neg$(**Bel** $\neg$S)

i $\models$ [A] **Bel** S

In the next case the outcome of A is uncertain at index i; after doing A, the agent will believe that S or believe that $\neg$S, though initially he believes neither.

Figure 2

That verifies

i $\models \neg$(**Bel** S) & $\neg$(**Bel** $\neg$S)

i $\models$ [A] (**Bel** S) $\vee$ (**Bel** $\neg$S)

In both of those examples, it seems reasonable to say that A has been informative. Is it possible to define when is an action informative? Yes, but it turns out to be more elegant to proceed indirectly. Do not just ask about an action "Is it informative?", but ask instead the two questions "Does it preserve information?" and "Does it preserve ignorance?". (Ignorance can also be described more clumsily as "lack of information".) Those questions can be answered "yes" and "no" in four combinations. One particular combination describes informative actions. In a table,

the action...	is information preserved?	is ignorance preserved?
...leaves information unchanged	yes	yes
...is informative	yes	no
...loses information	no	yes
...revises information	no	no

However, it is easier to start the discussion in terms of "information-gaining" and "information-losing" actions, and I shall do that.

Here is an example of an information-gaining action. The agent is playing the shell game. In front of him are three cups, upside down. He believes that under one of them is a bean, and he has to find where it is. He turns the left cup over. Either the bean is under it or it isn't. If it is, splendid. If it isn't, he still has more information about where the bean is. What has happened can be drawn like this:

Figure 3

The agent has then gained information. The important thing is that the action "splits" an information set into several smaller information sets. Indeed, my intuition (and Moore's, before me) is that that is what is crucial to being an informative action.

Below are informal diagrams of two ways an action A could be informative. The regions with thick outlines are entire information sets.

Figure 4

There may be some indices where A does not occur.

Figure 5

How can one characterize such actions? It seems that the essential thing is that they obey

$$(BEL\ ;\ \mathrm{A}) \supseteq (\mathrm{A}\ ;\ BEL)$$

That is, starting at any index, the indices accessible by first moving across a *BEL* arc and then across an action A arc are a superset of the indices accessible by first moving across an action A arc and then across a *BEL* arc.

Here is a diagram illustrating that. One can get from i to k by *BEL* ; A but not by A ; *BEL* . To make the diagram easy to draw, I have assumed that *BEL* is an equivalence relation, but that is just a convenience. The account in this paper places no constraint on **Bel** at all.

Figure 6

In the diagram about the shell game,

$$\mathrm{i} \models (\mathbf{Bel}\ ;\ \mathrm{Action}) \dot{\supseteq} (\mathrm{Action}\ ;\ \mathbf{Bel})$$

and in fact that is true at every other index in the diagram too.

Another reason for thinking that this is the right way to describe informative action is that it leads to persuasive rules about the monotonicity of information. Using the rule about inclusion of modals

$$\models \text{Modal}_1 \stackrel{\bullet}{\supseteq} \text{Modal}_2 \rightarrow [\text{Modal}_1]\ \text{S} \rightarrow [\text{Modal}_2]\ \text{S}$$

it follows that

$$\models (\mathbf{Bel}\ ;\ \text{A}) \stackrel{\bullet}{\supseteq} (\text{A}\ ;\ \mathbf{Bel}) \rightarrow \mathbf{Bel}\ [\text{A}]\ \text{S} \rightarrow [\text{A}]\ \mathbf{Bel}\ \text{S}$$

which says "If I believe S will be true after A, then after doing A I will still believe S". That applies to any sentence S. Suppose that it is also believed that A preserves the truth of S — that is, if S is true before A then S is still true after A. Then one can say that A preserves belief about S, because it is easy to show that

$$\models (\mathbf{Bel}\ ;\ \text{A}) \stackrel{\bullet}{\supseteq} (\text{A}\ ;\ \mathbf{Bel}) \rightarrow \\ \mathbf{Bel}\ (\text{S} \rightarrow [\text{A}]\ \text{S}) \rightarrow (\mathbf{Bel}\ \text{S}) \rightarrow [\text{A}]\ \mathbf{Bel}\ \text{S}$$

It is also possible for an action to lose information. Suppose the agent has three cups upside down in front of him. He believes the bean is under the left cup. He shuts his eyes, and someone else slides the cups about. As a result of his action (of waiting) the agent has lost information. What has happened can be drawn like this:

Figure 7

The important thing there is that doing A "inserts" the initial information set into a larger information set. Again, that seems to be what is crucial for being an information-losing action. Here are two informal diagrams of an information-losing action A.

Figure 8

Figure 9

What characterizes such cases is that they obey

$$(\mathit{BEL}\ ;\ \text{A}) \subseteq (\text{A}\ ;\ \mathit{BEL})$$

Here is a diagram to illustrate that. One can get from i to k by A ; *BEL* but not by *BEL* ; A. Again, *BEL* is shown as an equivalence relation; again, it need not be.

Figure 10

Such actions show a sort of monotonicity of ignorance. Again using the rule about inclusion of modals, it follows that

$$\models (\mathbf{Bel}\ ;\ \mathrm{A}) \dot{\subseteq} (\mathrm{A}\ ;\ \mathbf{Bel}) \rightarrow [\mathrm{A}]\ \mathbf{Bel}\ \mathrm{S} \rightarrow \mathbf{Bel}\ [\mathrm{A}]\ \mathrm{S}$$

which say that "If I believe S after doing A, then I already expect to believe S after doing A"; or, taking a contrapositive,

$$\models (\mathbf{Bel}\ ;\ \mathrm{A}) \dot{\subseteq} (\mathrm{A}\ ;\ \mathbf{Bel}) \rightarrow \neg(\mathbf{Bel}\ [\mathrm{A}]\ \mathrm{S}) \rightarrow \neg([\mathrm{A}]\ \mathbf{Bel}\ \mathrm{S})$$

which says "If I do not already believe S will be true after A, then I will not come to believe S as a result of doing A". Suppose that it is believed that A preserves the falsity of S; if S is true after A then S must have been true before A. (One might expect that to be formulated as "if S is false before A then S is still false after A". In fact that runs into problems if A is impossible.) Then one can say that A preserves lack of belief in S, because it is easy to show that

$$\models (\mathbf{Bel}\ ;\ \mathrm{A}) \dot{\subseteq} (\mathrm{A}\ ;\ \mathbf{Bel}) \rightarrow \\ \mathbf{Bel}\ (([\mathrm{A}]\ \mathrm{S}) \rightarrow \mathrm{S}) \rightarrow ([\mathrm{A}]\ \mathbf{Bel}\ \mathrm{S}) \rightarrow \mathbf{Bel}\ \mathrm{S}$$

I shall define an action A as being "information-preserving" at an index i iff

$$\mathrm{i} \models (\mathbf{Bel}\ ;\ \mathrm{A}) \dot{\supseteq} (\mathrm{A}\ ;\ \mathbf{Bel})$$

and as "ignorance-preserving" iff

$$\mathrm{i} \models (\mathbf{Bel}\ ;\ \mathrm{A}) \dot{\subseteq} (\mathrm{A}\ ;\ \mathbf{Bel})$$

I said that it was better to take classify actions as "information-preserving" and "ignorance-preserving" rather than as "information-gaining" and "information-losing". To see why that pays off, consider this: I am playing the shell game, and I start with three cups upside down in front of me. I do not know which covers the bean. I slide the left and the right cups into each other's place. The situation could be drawn as

Figure 11

I have changed the state of the world, but neither gained nor lost information. After my action, my information set is exactly the "image" under the action of my initial information set, neither more nor less. Since

$$\mathrm{i} \models (\mathbf{Bel}\ ;\ \mathrm{A}) \dot{\supseteq} (\mathrm{A}\ ;\ \mathbf{Bel})$$

$$\mathrm{i} \models (\mathbf{Bel}\ ;\ \mathrm{A}) \dot{\subseteq} (\mathrm{A}\ ;\ \mathbf{Bel})$$

A should (at i) be classified as both information-preserving and ignorance-preserving, which seems right. It therefore follows that both

$$\mathrm{i} \models [\mathrm{A}]\ \mathbf{Bel}\ \mathrm{S} \rightarrow \mathbf{Bel}\ [\mathrm{A}]\ \mathrm{S}$$

$$\mathrm{i} \models \mathbf{Bel}\ [\mathrm{A}]\ \mathrm{S} \rightarrow [\mathrm{A}]\ \mathbf{Bel}\ \mathrm{S}$$

so

$$i \models [A]\ \mathbf{Bel}\ S \equiv \mathbf{Bel}\ [A]\ S$$

which is exactly what one would expect. If I had had to classify actions in terms of information-gaining or information-losing, neither would apply in this case, and I could say nothing about it.

The last case is actions that are neither information-preserving or ignorance-preserving. Imagine in the shell game that I turn up all the cups and find that none of them covers the bean. That could be drawn as below. The actual index is i.

Figure 12

I have discovered that some of my beliefs were wrong, and now my information set contains some states of affairs that previously I thought impossible. The actual index i was not epistemically accessible from itself. The agent could not believe that he would end up at j even though he does. Neither of

$$i \models (\mathbf{Bel}\ ;\ A) \dot{\supseteq} (A\ ;\ \mathbf{Bel})$$

$$i \models (\mathbf{Bel}\ ;\ A) \dot{\subseteq} (A\ ;\ \mathbf{Bel})$$

are true, and nothing follows. Again, that reasonable; such actions lead to arbitrary and unpredictable revisions of belief.

The vital point about that last example was that the expected future information set and the information set that actually happened were unrelated. That the agent may have wrong belief, so that the actual index is not in its own information set, is perfectly compatible with this account. Here for instance is a diagram of an abstract case where an action is informative even though the beliefs about which it is informative may be wrong.

Figure 13

Characterizing actions that preserve knowledge in this way has been done by (Werner 89) and (Lehmann Kraus 86). Characterizing actions that preserve ignorance is I believe novel, and so is the idea of four cases of information change identified by the answers to two separate questions about information preservation.

2.1 The litmus example

Here is an example of how the rules given above make it possible to reason about informative action. I take a version of Moore's example of dipping litmus into a solution to see whether it is acid or not. This particular example does not go beyond Moore: it is here for purposes of comparison.

The immediately obvious application is wrong. It is certainly true that dipping the litmus is information-preserving, so

(**Bel** ; dip) $\dot{\supseteq}$ (dip ; **Bel**)

Hence

(**Bel** [dip] red) → ([dip] **Bel** red)

and, since there is no doubt about what litmus does,

invariably (acid → [dip] red)

so

Bel (acid → [dip] red)

then

(**Bel** acid) → ([dip] **Bel** red)

That is true, but not very interesting, since the antecedent is false whenever using litmus is worthwhile.

So why is litmus useful? Because after dipping the litmus, it is red iff only the solution is acid

invariably [dip] (acid ≡ red)

hence both of

Bel [dip] (acid ≡ red)

Bel [dip] (¬acid ≡ ¬red)

The next step is where the relation between action and belief matters. It is only because dipping is information-preserving that one can go on to infer

[dip] **Bel** (acid ≡ red)

[dip] **Bel** (¬acid ≡ ¬red)

Box modals distribute across equivalences, so

[dip] ((**Bel** acid) ≡ (**Bel** red))

[dip] ((**Bel** ¬acid) ≡ (**Bel** ¬red))

It is indubitable that one knows whether the litmus is red:

invariably ((**Bel** red) ≢ (**Bel** ¬red))

so that is true after any included modal:

[dip] ((**Bel** red) ≢ (**Bel** ¬red))

and, since one believes that the litmus's being red (or not) after dipping is equivalent to the solution's being acid (or not),

[dip] ((**Bel** acid) $\not\equiv$ (**Bel** $\neg$acid))

The rest of this section is here just to calm any feeling that there was a cheat in that account; that it started with

invariably [dip] (acid $\equiv$ red)

when there is a deeper explanation of our trust in litmus, based on

invariably (acid $\rightarrow$ [dip] red)

There is, but it has to do with frame axioms — statements about what does not change when an action is performed — rather than with belief. Such axioms are

invariably (acid $\rightarrow$ [dip] acid)

invariably ($\neg$acid $\rightarrow$ [dip] $\neg$acid)

Furthermore, litmus would be a useless test if sometimes it turned red even for non-acids. Fortunately

invariably ($\neg$acid $\rightarrow$ [dip] $\neg$red)

It is logically necessary that

invariably (acid $\vee$ $\neg$acid)

Suppose that the solution is acid. Then, after dipping, it is still acid, and the litmus will be red. So, in that case, after dipping, redness and acidity are equivalent. Similarly if the solution is not acid. So

invariably [dip] (acid $\equiv$ red)

No argument about belief was needed to get that.

2.2 Preserving information about a single sentence

It is also possible to characterize those actions which preserve information about a particular sentence S rather than in general. One possibility would be to say that they were those actions that obeyed

(**Bel** S) $\rightarrow$ [A] **Bel** S

But it would be nicer to read a rule off a semantic property. The thing to demand is that action A is information-preserving at least when restricted to indices where S is true — that is, that A satisfies

$(BEL\ ;\ [\![\ \mathrm{S?}\ ]\!]\ ;\ \mathrm{A}) \supseteq (\mathrm{A}\ ;\ BEL)$

Remember that $[\![\text{S?}]\!]$ is the relation that links each index where S is true to itself and links nothing else. (The arguments M, g, are obvious from context.)

$$[\![\text{S?}]\!]\ \text{M g} \triangleq \{ \langle \text{ii} \rangle \mid \text{M g i} \models \text{S} \}$$

Then one can say "If action A is informative about S, and I believe that S is sufficient grounds for A achieving U, then after A I will believe U". Of course U can be S. Formally,

$$\models (\mathbf{Bel}\ ;\ \text{S?}\ ;\ \text{A}) \dot{\supseteq} (\text{A}\ ;\ \mathbf{Bel}) \rightarrow \mathbf{Bel}\ (\text{S} \rightarrow [\text{A}]\ \text{U}) \rightarrow [\text{A}]\ \mathbf{Bel}\ \text{U}$$

That follows at once from

$$\models \text{Modal}_1 \dot{\supseteq} \text{Modal}_2 \rightarrow [\text{Modal}_1]\ \text{S} \rightarrow [\text{Modal}_2]\ \text{S}$$
$$\models [\text{S?}]\ \text{T} \equiv (\text{S} \rightarrow \text{T})$$

If one takes S to be constant truth, then the information-preserving actions discussed earlier arise as a special case.

To go back to the litmus example, one thing that is *not* true is that dipping is "informative" about acidity. It is true that — it is an instance of a theorem that —

$$(\mathbf{Bel}\ ;\ \text{acid?}\ ;\ \text{dip}) \dot{\supseteq} (\text{dip}\ ;\ \mathbf{Bel}) \rightarrow$$
$$\mathbf{Bel}\ (\text{acid} \rightarrow [\text{dip}]\ \text{red}) \rightarrow \mathbf{Bel}\ [\text{dip}]\ \text{red}$$

which might start an interesting deduction; but the antecedent is false. What the antecedent says is that all states of affairs reachable by dipping the litmus arise from a state of affairs where the solution is acid; which is not so.

One can also define actions that preserve ignorance about a sentence S; but I can't see why one would want to.

3 Finding that, finding whether, finding a, and evaluating

3.1 Overview

This section starts with an overview of abstract actions on belief, and then describes in detail how to construct them.

It is possible to characterize actions as being informative; it is also possible to say "Action A is a suitable one to do to find whether S" by asserting

$$[\text{A}]\ (\mathbf{Bel}\ \text{S}) \vee (\mathbf{Bel}\ \neg\text{S})$$

It would be nice to construct the action of coming to know about S, **find whether S** say, as an action in its own right, which could be incorporated in plans just like any action on the world. It should obey something like

$$\models \ldots \rightarrow [\textbf{find whether } \mathrm{S}]\ (\textbf{Bel } \mathrm{S}) \vee (\textbf{Bel } \neg \mathrm{S})$$

It is possible to describe such an action, but again it turns out to be better to proceed indirectly. I shall first introduce a **find that** S action, which will obey something like

$$\models \ldots \rightarrow [\textbf{find that S}]\ \textbf{Bel } \mathrm{S}$$

The antecedents ...turn out to be pretty weak, so isn't that rule wildly too strong? Can one not use it to persuade oneself of anything one likes, whenever one wants to, so that it is more appropriate as a representation of psychosis than of discovery? No, because what happens is that one won't be able to do it unless S is something that one can sensibly come to believe. Then one can define **find whether** S as a compound action, the union of trying to execute instances of both accepting that S and accepting that ¬S, and then rely on the fact that one of them must succeed.

An action such as **find that** S can be a mental just as well as a physical action. What matters about it is its effect on belief. An appropriate change can be brought about about by inference or reflection just as well as by physical tests and observations. If other actions, such as **find whether** S, are defined in terms of **find that** S, then they can be seen as potentially mental too. Plans that include them are plans to think as well as to do.

In that sketch of finding whether S is true or not, the inner structure of S was ignored; the action on belief was "propositional". A similar action on belief is finding an object that has a certain property ϕ; such an action is "first-order". It ought to be possible to describe such an action, **find a** ϕ say, which obeys something like

$$\models \ldots \rightarrow [\textbf{find a } \phi]\ \exists \mathrm{x}.\ \textbf{Bel } \phi(\mathrm{x})$$

After the action, there will be some fixed object that is believed to be ϕ, regardless of what term is used to describe it. That outcome is quite different from the much less interesting

$$\textbf{Bel } \exists \mathrm{x}.\ \phi(\mathrm{x})$$

Given that the first-order analogue of ∨ often turns out to be ∃, the action of **find a** ϕ is the obvious analogue of the action of **find whether** S. **Find a** ϕ can be seen as the union of many informative actions of the form **find that** $\phi(\mathrm{x})$ where x can be any of the objects in the domain.

A related sort of action on belief is evaluation. Suppose I am faced with a complex arithmetic expression, such as $723 * 114$. If I accept the theory of arithmetic, I must suppose that that term has a denotation; but I do not know what it is. But by thinking about it, I can work out that it is in fact 82422. The action on belief of evaluating — starting with a term whose denotation is uncertain and coming to a state where its denotation is fixed — can be seen as **find a** ϕ when ϕ has the form

λy.(y = Term)

Its result will be something of the form

$\exists$y. **Bel** (y = Term)

3.2 Finding that

Here are the details about **find that** S. It is essentially an auxiliary notion, used to define more important notions later. It must be a relation on indices like any other action. What would it look like? Each information set it leads to must be the image of a subset of the original information set throughout which S is true. It is not necessary that an index and its image agree about S. It is possible that the truth of S change while one finds out about it. Its simplest form would be like this:

Figure 14

Notice that indices where S is false have no image.

The next example would also be allowable; an information set is "split" into more than two parts, but each part contains only images of indices where S was true.

Figure 15

And so would this, where at some of the indices in the information set where the action could occur it does not.

Figure 16

And so would this, where the action is non-deterministic.

Figure 17

And so would this, where admittedly the agent is mistaken, but all the indices he comes to accept are images of indices where S was true.

Figure 18

In all of those, what matters is that A satisfies

$$(BEL\ ;\ [\![\ \mathrm{S?}\ ]\!]\ ;\ \mathrm{A}) \supseteq (\mathrm{A}\ ;\ BEL)$$

That criterion has already appeared as the test for whether an action A preserves information about S. To define the action of finding that S — the least specific action of that sort — take the union of all actions that preserve information about S. Define the modal **find that** S by

$$[\![\ \textbf{find that}\ \mathrm{S}\ ]\!]\ \mathrm{M}\ \mathrm{g} \triangleq \bigcup \mathrm{X}.\ ((BEL\ ;\ [\![\ \mathrm{S?}\ ]\!]\ ;\ \mathrm{X}) \supseteq (\mathrm{X}\ ;\ BEL))$$

Now I must show that that action has the property it should — that

$$\models \mathbf{Bel}\ (S \rightarrow [\mathbf{find\ that}\ S]\ U) \rightarrow [\mathbf{find\ that}\ S]\ \mathbf{Bel}\ U$$

(Remember U can be S.) Of course, as stated earlier,

$$\models (\mathbf{Bel}\ ;\ S?\ ;\ A) \stackrel{\bullet}{\supseteq} (A\ ;\ \mathbf{Bel}) \rightarrow \mathbf{Bel}\ (S \rightarrow [A]\ U) \rightarrow [A]\ \mathbf{Bel}\ U$$

but is it true that the antecedent is true for **find that** S ? — that is,

$$\models (\mathbf{Bel}\ ;\ S?\ ;\ \mathbf{find\ that}\ S) \stackrel{\bullet}{\supseteq} (\mathbf{find\ that}\ S\ ;\ \mathbf{Bel})$$

It would be nice if, given a property P, the union of all the sets with that property also had the property P — that is, if

$$P(\bigcup X.\ P(X))$$

Then one could take P(X) to be the property

$$(BEL\ ;\ [\![\ S?\ ]\!]\ ;\ X) \supseteq (X\ ;\ BEL)$$

and the proof would be immediate. Unfortunately it is not so in general: take P(X) to be "X is a singleton". Fortunately, it is true in this case.

$$P(\bigcup X.P(X)) \text{ where } P(X) \text{ is } A;X \supseteq X;B$$

Proof: see appendix A.2

To show this action is not too powerful, suppose that the actual index is i, and consider what happens if the agent believes that S is false before doing **find that** S. The initial information set looks like this

Figure 19

so the part of the relation $BEL\ ;\ [\![\ S?\ ]\!]$ "local" to i — that is, its restriction to i,

$$\{\ \langle ij\rangle \mid \langle ij\rangle \in BEL\ ;\ [\![\ S?\ ]\!]\ \}$$

—, must be the empty relation, so if

$$(BEL\ ;\ [\![\ S?\ ]\!]\ ;\ X) \supseteq (X\ ;\ BEL)$$

then either the part of X local to i, or the part of *BEL* local to i, must be empty. So **find that** S cannot occur without the information set afterwards being empty.

It is not necessary to adopt and verify a rule to that effect. The right result follows in the object language.

$$\models (\mathbf{Bel}\ \neg S) \rightarrow [\mathbf{find\ that}\ S]\ \mathbf{Bel}\ \#$$

Proof: see appendix A.3

That is, if I try to find that S when I believe ¬S, I end up in an absurd situation when my information set is empty, and so I believe everything. If, as usually happens, that is ruled out as impossible, so that belief obeys the axiom often called "D"

$$\models (\mathbf{Bel}\ \#) \rightarrow \#$$

then I can go on to conclude that

[**find that** S] #

— that is, **find that** S is impossible — which seems right.

3.3 Finding whether

Now to construct the action of finding whether S is true or not, by combining finding that S and finding that ¬S. Since a **find that** S action selects some subset of the indices where S is true and makes its image one possible information set after the action is over, the **find whether** S action is like a whole bundle of those actions in parallel, separating the original information set into many information sets. Each will be the image of a set of indices in which S is true throughout or false throughout. I picture it as

Figure 20

Define

$$[\![\ \mathbf{find\ whether}\ \mathrm{S}\ ]\!] \triangleq [\![\ (\mathbf{find\ that}\ \mathrm{S}) \cup (\mathbf{find\ that}\ \neg\mathrm{S})\ ]\!]$$

(Notice that **find whether** S is the same as **find whether** ¬S.) Now it is easy to show that

$$\models \mathbf{Bel}\ (\mathrm{S} \rightarrow [\mathbf{find\ that}\ \mathrm{S}]\ \mathrm{U})\ \&\ \mathbf{Bel}\ (\neg\mathrm{S} \rightarrow [\mathbf{find\ that}\ \neg\mathrm{S}]\ \mathrm{V}) \rightarrow [\mathbf{find\ whether}\ \mathrm{S}]\ (\mathbf{Bel}\ \mathrm{U}) \vee (\mathbf{Bel}\ \mathrm{V})$$

(Both U and V can be S.)

Proof: Immediate from definition of **find whether** S and

$$\models [\mathrm{A}]\ \mathrm{S}\ \&\ [\mathrm{B}]\ \mathrm{S} \equiv [\mathrm{A} \cup \mathrm{B}]\ \mathrm{S}$$

End proof

3.4 Finding a

I suggested that there should be a representation of finding an object that has a certain property ϕ. It was to be a "first-order" analogue of the "propositional" action of deciding whether S was true or not, and this is what it should achieve: "If I believe that all objects with the property ϕ before the action will still have the property ψ afterwards, then after the action there will be an object that I believe to be ψ". ψ can of course be ϕ. Putting it formally,

$\models$ [**find a** ϕ] $\exists$x. **Bel** ϕ(x)

For instance, deciding on an attractive hardy plant to put in a garden would be

find a λx. (plant(x) & hardy(x) & attractive(x))

What would such an action denote? Its simplest form is more complex than in the **find whether** case, where at any index, S is exactly one of true and false, because at each index there may be one, or many, or no, objects that satisfy ϕ. For any object d, there is a subset of the initial information set containing exactly those indices where ϕ is true of d. The vital requirement on the **find a** ϕ action is that each information set into which the initial information set is "split" should spring from within a single one of those subsets. The complication is that those subsets need be neither exclusive nor exhaustive. Again, it does not follow that just because ϕ is true of d at an index, it is still true at the images of that index, but if that is so, then in the image of each subset there is at least one object which is ϕ at all the indices, and which can therefore be said to be believed to be ϕ.

As when **find whether** S was constructed out of **find that** S and **find that** $\neg$S, so **find a** ϕ is best constructed out of something simpler. That something simpler turns out to be **find that** S again, but where S now has the form ϕ(x). Remember that for a fixed variable assignment, "x" is the name of an object in the domain. What **find that** ϕ(x) does is form images of subsets of the original information set throughout which ϕ(x) is true. Here is a diagram of such a case.

Figure 21

Notice that indices where x is not ϕ have no image. Here is a case where the image is fragmented more than it "need" be.

Figure 22

Suppose that in the initial information set there are several subsets of indices, where each subset is exactly the indices where some particular object has the property ϕ. It could look like this.

Figure 23

There is a possibility of **find a** ϕ being, in a way, "non-deterministic". If there is more than one object that is ϕ, presumably any one of them might be found. How does that manifest?

Find a ϕ must form new information sets, ensuring that each one is rooted in exactly one of the original subsets. There are many ways that can happen, though only one of them will occur in a given frame. The "non-determinism" is at the level of frames rather than indices. What exactly **find a** ϕ denotes depends on the "pre-existing" *BEL* relation in that frame, in terms of which **find that** ϕ(x) is defined. For instance, the image of the information set in the previous figure under **find a** ϕ could, in different frames, be any of these:

Figure 24

Define

$$[\![\ \textbf{find a}\ \phi\ ]\!] \triangleq [\![\ \cup x.\ \textbf{find that}\ \phi(x)\ ]\!]$$

That definition means that the informally-stated rule proposed at the start of this section is true.

$$\models \forall x.\ \textbf{Bel}\ (\phi(x) \rightarrow [\textbf{find that}\ \phi(x)]\ \psi(x)) \rightarrow [\textbf{find a}\ \phi]\ \exists x.\ \textbf{Bel}\ \psi(x)$$

Proof: Consider the rule that **find that** S obeys when S contains free variables:

$$\models \textbf{Bel}\ (\phi(x) \rightarrow [\textbf{find that}\ \phi(x)]\ \psi(x)) \rightarrow [\textbf{find that}\ \phi(x)]\ \textbf{Bel}\ \psi(x)$$

The rest of the proof is almost immediate from definition of **find a** ϕ and

$$\models \forall x.([A]\ S) \equiv [\cup x.\ A]\ S \qquad \text{if x is not free in S}$$

End Proof

What happens if one tries to find something that is ϕ when there is isn't any such thing? Something ought to go wrong; fortunately, it does.

$$\models (\textbf{Bel}\ \neg\exists x.\ \phi(x)) \rightarrow [\textbf{find a}\ \phi]\ \textbf{Bel}\ \#$$

Proof: see appendix A.4

As before, if belief obeys axiom D, that shows that **find a** ϕ is impossible which seems right.

3.5 Evaluating

One particular sort of action on belief is evaluating a complex term such as "Tom's eldest cousin's mother's father" to find out what object it denotes. For a while I thought that this was a sort of mental action I could not handle until I saw that it was nothing more than finding the (unique) object satisfying the predicate

$$\lambda\ y.(y = \text{father(mother(eldest-cousin(Tom))))}$$

In general, and suppressing irrelevant preconditions, the result of finding an object that satisfies

$$\lambda y.(y = \text{Term})$$

is

$$\ldots. \rightarrow [\textbf{find a}\ \lambda y.(y = \text{Term})]\ \exists x.\ \textbf{Bel}\ (\lambda y.(y = \text{Term}))(x)$$

which simplifies to

$$\ldots . \rightarrow [\textbf{find a } \lambda y.(y = \text{Term})]\ \exists x.\ \textbf{Bel}\ (x = \text{Term})$$

which is a claim that the agent knows what Term denotes. One can define

$$[\![\ \textbf{evaluate}\ \text{Term}\ ]\!] \triangleq [\![\ \cup x.\ \textbf{find that}\ x = \text{Term}\ ]\!]$$

which obeys

$$\models [\textbf{evaluate}\ \text{Term}]\ \exists x.\ \textbf{Bel}\ (x = \text{Term})$$

4 Action on belief relative to a theory

4.1 Overview

One sort of action on belief is to infer information from a theory. This section discusses how that can be done without having to suppose that the agent is logically omniscient about the consequences of the theory.

There is no doubt that deduction and evaluation are actions; they take time and effort, they can be useful, if I attempt them I may succeed or fail, and so on. They are different from ordinary actions in that what they change is not the physical world, but my beliefs. I want to represent them in dynamic and epistemic logic, just as for actions on the world; but that runs straight into the problem of logical omniscience. In this context logical omniscience would make all action on belief pointless.

Suppose I believe some theory of mechanics — call it Mech; that is, the actual index is i and

$$i \models \textbf{Bel}\ \text{Mech}$$

(For the sake of the example I assume that both my theory of mechanics and my knowledge of the initial state of affairs are complete; of course in reality I would never have that.) Suppose I want to calculate whether a certain billiard ball on a certain table will go into a certain pocket — call that fact GoesIn. Suppose that Mech does in fact give an answer about what happens to the ball; that is $\models$ Mech $\rightarrow$ GoesIn or $\models$ Mech $\rightarrow \neg$GoesIn Now assume $\models$ Mech $\rightarrow$ GoesIn. Since i $\models$ **Bel** Mech, then i $\models$ **Bel** GoesIn. Similarly, if $\models$ Mech $\rightarrow \neg$GoesIn, then i $\models$ **Bel** $\neg$GoesIn. So I am not in doubt about the truth of GoesIn. This is general. If I believe a theory T, and that theory is able to answer a question, it seems that I already know the answer to the question, and that the effort of deduction must be futile.

The dilemma here is sharp: if a standard Hintikka-style epistemic logic is used, and if a theory is believed, the consequences of the theory must be believed too. If I wish to avoid logical omniscience, I must either forgo Hintikka-style logic or deny that theory is believed. The first approach has been tried (eg in the

line of work started by Levesque around 1984 and appearing in (Levesque 1987)) and has proved popular. But if it is pursued, belief and action are no longer represented in the same way or in the same sort of modal logic. For instance, (Levesque 1987) associates valuations, not first-order models, with indices. That is much more natural for a logic of belief than of action. Furthermore, such a revision of epistemic logic is going to make claims about what the believer always believes, not what he believes after thought and calculation, which is what a psychological account would demand.

If I want to keep the logics the same, I must take the other branch. My feeling is that a logicist account of imperfect reasoning should not offer a language and semantics in which the sound inferences capture imperfect but psychologically correct reasoning, but instead a set of assumptions from which results of the imperfect reasoning follow in accordance with a standard language and semantics.

Thus, to avoid an agent's logical omniscience about the consequences of a theory, I deny that the agent believes the theory — that is, I deny that it is true throughout its information set. Nevertheless, the theory T will still be true at some indices. Suppose that the actual index is i. Call the subset of the information set where T is true, that is, the set

$$\{ \mathrm{j} \mid \langle \mathrm{ij} \rangle \in \mathit{BEL} \ \& \ \mathrm{j} \models \mathrm{T} \}$$

the "core". Finding out whether Mech says GoesIn is true or not will be seeing what Mech says about GoesIn, and rejecting all those indices that say something different. The remainder must include the core, but will usually also include other indices which happen to be right about GoesIn but wrong about some other sentence which has not yet been thought about. This sort of finding out is a process of shrinking the information set so that it always includes the core but excludes indices that have turned out to be mistaken. This diagram shows the intuition.

Figure 25

The intended application is to reflexive planning systems that can create plans with "make a plan" steps in the plan itself. Such steps need a semantics, which they will get by being seen as **find a** actions — for instance, "**find an** action. Goal is true after action". If planners are logically omniscient, such steps are possible but pointless since the planner already knows a suitable plan. If however it makes sense to say that it does not know all the consequences of its theory of change and action, they are not pointless. This section is an attempt to show that it does make sense.

4.2 Details

What I am trying to represent is trying to find out that S, while basing one's enquiry on the theory T; trying to deduce S from T would be such an action.

If I call that action "A", then I can draw it as

Figure 26

What happens if one tries that when T asserts that S is false? It sounds like attempting something impossible. Impossible actions denote the empty relation. If one believes that T entails ¬S, the indices where one believes T are a subset of the indices where one believes ¬S. What should then happen can be suggested by

Figure 27

The action A is empty and the original information set has no image.

What should happen if the theory T does not decide S? That can be drawn as

Figure 28

What happens depends on what one takes the action on belief to be. If it is a sort of checking of S against T, to see if S is compatible with T, then the action should succeed; I shall come back to it. If on the other hand it is like deduction, then the action should again be impossible (empty) because clearly one does not believe that T entails S.

To represent seeing if S is entailed by T, I propose something very like the action of finding that S, but further restricted by the requirement that the extension of S (at least at credible indices) be a superset of the extension of T (at least at credible indices). That further requirement is enforced by Q in this definition.

$$[\![\textbf{find that } S \textbf{ accepting } T]\!]\ M\ g \triangleq \bigcup X.\ (P(X)\ \&\ Q)$$

where P(X) is $(BEL\ ;\ [\![S?]\!]\ ;\ X) \supseteq (X\ ;\ BEL)$
and Q is $(BEL\ ;\ [\![S?]\!]) \supseteq (BEL\ ;\ [\![T?]\!])$

It is easy to verify that Q is true iff at all indices the agent believes that T implies S; that is, iff

for some i, M i $\models$ **invariably Bel** (T $\rightarrow$ S)

I interpret that as a demand that S follow from T with perhaps an appeal to facts that are invariably even if not logically true, but without relying on any merely contingent facts, true at some indices but false at others. In that case, **find that** S **accepting** T denotes the same relation as **find that** S. If on the other hand Q is false, so that the action of deduction of S from T is impossible, then **find that** S **accepting** T is empty; as it should be. It is easy to verify that

$$\models \textbf{invariably Bel}\ (T \rightarrow S) \rightarrow \textbf{find that}\ S\ \textbf{accepting}\ T \stackrel{\bullet}{=} \textbf{find that}\ S$$

$$\models \neg(\textbf{invariably Bel}\ (T \rightarrow S)) \rightarrow \textbf{find that}\ S\ \textbf{accepting}\ T \stackrel{\bullet}{=} \textbf{abort}$$

(In the special case of T being constant truth, so that the antecedent is

invariably Bel S

one is looking for S following from no special assumptions at all.)

Those rules allow let one prove that **find that** S **accepting** T and **find that** S both make this valid

$$\models \textbf{Bel}\ (S \rightarrow [X]\ U) \rightarrow [X]\ \textbf{Bel}\ U$$

when substituted for X. The difference between them is when they are impossible — that is, when they are equal to **abort**.

Just believing $T \rightarrow S$ is not enough to make **find that** S **accepting** T and **find that** S the same; it has to be "invariably" believed true. This is false:

$$\models \textbf{Bel}\ (T \rightarrow S) \rightarrow \textbf{find that}\ S\ \textbf{accepting}\ T \stackrel{\bullet}{=} \textbf{find that}\ S$$

The action of checking of S against T, to see if S is compatible with T, is a sort of dual to the action of deduction. It can be done when the extensions of T and S overlap, as suggested in

Figure 29

but should fail when they don't, as here

Figure 30

When it succeeds, the core is not preserved entirely, but some of it does survive into the image. Such an action can be defined by

$$[\![\ \textbf{find that}\ S\ \textbf{not rejecting}\ T\ ]\!]\ M\ g \stackrel{\triangle}{=} \bigcup X.\ (P(X)\ \&\ Q')$$

where P(X) is as before $(BEL\ ;\ [\![\ S?\ ]\!]\ ;\ X) \supseteq (X\ ;\ BEL)$
and Q′ is $\neg((BEL\ ;\ [\![\ S?\ ]\!]\ ) \subseteq (BEL\ ;\ [\![\ \neg T?\ ]\!]\ ))$

Then Q′ is true iff at all the indices the agent believes that T is compatible with S; that is, iff

for some i, M i $\models \neg\textbf{invariably Bel}\ (T \rightarrow \neg S)$

Then

$$\models \textbf{invariably Bel}\ (T \rightarrow \neg S) \rightarrow \textbf{find that}\ S\ \textbf{not rejecting}\ T \stackrel{\bullet}{=} \textbf{abort}$$

$$\models \neg(\textbf{invariably Bel}\ (T \rightarrow \neg S)) \rightarrow \textbf{find that}\ S\ \textbf{not rejecting}\ T \stackrel{\bullet}{=} \textbf{find that}\ S$$

As before, **find that** S **not rejecting** T behaves exactly like **find that** S — for instance,

$$\models \textbf{Bel}\ (S \rightarrow [X]\ U) \rightarrow [X]\ \textbf{Bel}\ U$$

where X is **find that** S **not rejecting** T — except that it is impossible when it needs to be.

4.3 Derived actions on belief

Actions of **find whether** and so on were defined in terms of **find that**. They all have analogues defined in terms of **find that** ... **accepting** .... The definitions are merely

$$[\![\ \textbf{find whether}\ S\ \textbf{accepting}\ T\ ]\!] \triangleq [\![\ \textbf{find that}\ S\ \textbf{accepting}\ T \cup \textbf{find that}\ \neg S\ \textbf{accepting}\ T\ ]\!]$$

$$[\![\ \textbf{find a}\ \phi\ \textbf{accepting}\ T\ ]\!] \triangleq [\![\ \cup x.\ \textbf{find that}\ \phi(x)\ \textbf{accepting}\ T\ ]\!]$$

$$[\![\ \textbf{evaluate}\ \text{Term}\ \textbf{accepting}\ T\ ]\!] \triangleq [\![\ \cup x.\ \textbf{find that Bel}\ (x = \text{Term})\ \textbf{accepting}\ T\ ]\!]$$

For instance, solving the problem in arithmetic above would be doing

evaluate $723 * 114$ accepting Arithmetic

There are similar analogues with "not rejecting T".

5 Related work

Related work might be about any of several closely-related topics. It might be about:

- How actions change belief

  > "If I read this telephone directory, I will know Mary's telephone number."

- How one can therefore plan to change beliefs

"I will read this telephone directory, in order to know Mary's telephone number."

- How belief enables action

  "If I know Mary's telephone number, I will be able to phone her."

- How one can therefore plan to enable action

  "I will read this telephone directory, in order to know Mary's telephone number, in order to be able to phone her."

This paper is about how actions change belief. It sounds as if there ought to be a great deal of related work, but I have been able to find much less than I would have expected. Instead, there is a lot of work on how belief enables action. Explaining that is indeed my long-term aim, but it is not the point of the present paper. Examples of work on the effect of belief on action are (Morgenstern 87), (Werner 89), (Werner 90), (Singh 91). It is interesting but (here) irrelevant and I shall not review it.

There are two groups of strictly related work: work explicitly on the topic, which appears to be tiny, and work which shares a metaphysics and some interest with what I have discussed here, but which has different aims.

Work in the first group start with Moore. He, in his thesis and later revisions (Moore 79, Moore 85), was the first to relate action and knowledge by looking at possible worlds connected by both action relations and epistemic accessibility, and to explain gain in information by the shrinking of the set of credible possible worlds. I have adopted that idea entire, but have, I believe, taken it further. Furthermore, the idea of actions characterized purely by their effect on belief is new.

Moore also presented his theory in what one usually sees as the meta-theory of a modal language; it quantifies over possible worlds, and uses various devices to express the denotation of syntactic objects. I have worked inside a modal logic, which, regardless of other merits and failings, is undoubtedly briefer.

Konolige (80) treats the same problem; but as he says, his main interest is his representation of belief, the "syntactic approach". He presents a theory in which the expressions of nested languages are first-class objects and uses that to express belief and change. His main interest is not to supplant or amend Moore's account, but to replicate it; as he says, to "show that the syntactic approach, when integrated with a situation calculus description of actions, can adequately formalize Moore's criteria for the interaction of knowledge and belief".

Shoham (Shoham 89) presents a very attractive synthesis of representations of time, change and belief. Our ignorance leaves us stranded, not just at one of a set of possible worlds, but at one of a whole rope of possible histories, each of which is a trajectory through state space as time passes. As we learn more,

the set of histories we might be on shrinks, and as we forget, the set enlarges. I find this a very attractive and fruitful view, but in the interests of parsimony and generality the paper considers possible worlds rather than possible histories. Pelavin (Pelavin 88) takes a similar view.

Related work in the second group (similar in spirit but not in aims) includes the situated automata of Rosenschein (85, 86) and the work of Halpern and Fagin (89). In both of those, what a system knows is what is true in all the states that it could be in which are compatible with the state that it actually is in. That is of course the standard definition of knowing from epistemic logic. What is new is that both approaches give a detailed and well-motivated account of the particular epistemic accessibility relation appropriate for the sort of agent or system that they are interested in — a perceiving robot, or a single agent in a distributed system. Once one has a complex and changing agent, it is very natural to start thinking about how what is believed changes with what is done.

6 References

Goldblatt, Robert: 1987
Logics of time and computation
Center for the study of language and information lecture notes #7
ISBN 0-937073-12-1

Hintikka, J: 1962
Knowledge and belief
Cornell UP: Ithaca, NY

Halpern, Joseph Y; Fagin, R: 1989
Modelling knowledge and action in distributed systems
Distributed computing 3 (1989) 159-177

Konolige K: 1980
A first-order formalization of knowledge and belief for a multi-agent planning system
in: Hayes JE, Michie D, Pao Y-Hin (eds)
Machine Intelligence 10. Ellis Horwood

Lehmann D, Kraus S: 1986
Knowledge, belief and time
Lecture Notes in Computer Science 226 186-195

Levesque H: 1987
All I know: an abridged report
Natl. Conf. on AI '87 (AAAI-87) 426-431. Amer. Assoc. for AI.

Moore, RC: 1979
Reasoning about knowledge and action
PhD thesis, Dept EECS, MIT (Feb 1979).

Moore, Robert: 1985
A formal theory of knowledge and action
in: Hobbs JR, Moore RC: 1985

Formal theories of the commonsense world
Norwood, NJ: Ablex publishing corp.
Morgenstern, Leora: 1987
Knowledge preconditions for actions and plans
Intl. Joint Conf. on AI '87 867-874
Pelavin, R: 1988
A formal approach to planning concurrent actions and external events
PhD, (also TR 254), Dept Computer Science, Rochester USA
Pratt, Vaughan R: 1976
Semantical considerations on Floyd-Hoare logic
IEEE symposium on foundations of computer science 17 (1976) 109-121
Rosenschein, Stanley J: 1985
Formal theories of knowledge in AI and robotics
New generation computing 3(4) (1985) 345-357
Rosenschein, Stanley J: 1986
The synthesis of digital machines with provable epistemic properties
in: Halpern, J (ed): 1986
Proc. Conf. on Theoretical Aspects of Reasoning 1986, 83-98. Morgan Kaufman
Shoham Y: 1989
Time for action: on the relation between time, knowledge and action
Intl. Joint Conf. on AI '89 954-959
Singh, Munindar: 1991
A logic of situated know-how
Natl. Conf. on AI '91 (AAAI-91) 343-348. Amer. Assoc. for AI.
Steel SWD: 1991
Sound substitution into modal contexts
Proc. Eighth Conference of SSAISB, Springer-Verlag, 1991
Werner, Eric: 1989
Modal Logic of Games.
WISBER Report Nr. B48, University of Saarbrücken
Werner, Eric: 1990
A unified view of information, intention and ability.
Second Euro. workshop on modelizing autonomous agents and multi-agent worlds.
(St Quentin en Yvelines; Aug 1990; ONERA) 69-83

A Proofs

A.1

$\models \forall x.([A]\ S) \equiv [\cup x.\ A]\ S$ if x is not free in S

Proof:

Direction $\rightarrow$

Assume that M g i $\models \forall x.([A]\ S)$,
so for any d $\in$ Dom, M g[x:=d] i $\models$ [A] S,
so if $\langle ij \rangle \in [\![A]\!]$ M g[x:=d] then M g[x:=d] j $\models$ S.
Assume that $\langle ij \rangle \in [\![\cup x.\ A]\!]$ M g,
so $\langle ij \rangle \in \bigcup d \in Dom.\ [\![A]\!]$ M g[x:=d]
so for some d $\in$ Dom, $\langle ij \rangle \in [\![A]\!]$ M g[x:=d]
Combining those, M g[x:=d] j $\models$ S.
If x is not free in S, that is equivalent to M g j $\models$ S.
Discharging the second assumption gives
if $\langle ij \rangle \in [\![\cup x.\ A]\!]$ M g then M g j $\models$ S
so M g i $\models [\cup x.\ A]$ S
Direction $\leftarrow$
Assume that M g i $\models [\cup x.A]$ S
Assume that is some d in Dom such that $\langle ij \rangle \in [\![A]\!]$ M g[x:=d]
so $\exists$d $\in$ Dom. $\langle ij \rangle \in [\![A]\!]$ M g[x:=d]
so $\langle ij \rangle \in \bigcup$ d $\in$ Dom. $[\![A]\!]$ M g[x:=d]
so $\langle ij \rangle \in [\![\cup x.\ A]\!]$ M g
Putting those together, M g j $\models$ S
and, if x is not free in S, M g[x:=d] j $\models$ S.
Discharging the second assumption gives
if $\langle ij \rangle \in [\![A]\!]$ M g[x:=d] then M g[x:=d] j $\models$ S
so M g[x:=d] i $\models$ [A] S.
Discharging the assumption that d $\in$ Dom, one gets
$\forall$d $\in$ Dom (M g[x:=d] i $\models$ [A] S)
so M g i $\models \forall x$ ([A] S)
End Proof

A.2

$P(\bigcup X.P(X))$ where P(X) is A ; X $\supseteq$ X ; B

Proof:
$P(\bigcup X.P(X))$ iff
A ; $(\bigcup X.P(X)) \supseteq (\bigcup X.P(X))$; B iff
$\langle ik \rangle \in (\bigcup X.P(X))$; B $\rightarrow \langle ik \rangle \in$ A ; $(\bigcup X.P(X))$
Now to show that that implication is true.
$\langle ik \rangle \in (\bigcup X.P(X))$; B iff
$\exists j.\ \langle ij \rangle \in \bigcup X.P(X)$ & $\langle jk \rangle \in$ B iff
$\exists j.\ (\exists X.\ P(X)$ & $\langle ij \rangle \in X)$ & $\langle jk \rangle \in$ B iff
$\exists X.\ P(X)$ & $\exists j.\ (\langle ij \rangle \in X$ & $\langle jk \rangle \in B)$ iff
$\exists X.\ P(X)$ & $\langle ik \rangle \in X$; B iff
$\exists X.$ A ; X $\supseteq$ X ; B & $\langle ik \rangle \in X$; B so
$\exists X.$ A ; X $\supseteq$ X ; B & $\langle ik \rangle \in$ A ; X iff (same, only backwards)
$\langle ik \rangle \in$ A ; $(\bigcup X.P(X))$

End Proof

A.3

$\models$ (**Bel** $\neg$S) $\rightarrow$ [**find that** S] **Bel** #

Proof:

1	**Bel** $\neg$S	assumption
2	$\neg$S	from 1
3	S	assumption
4	#	from 2,3
5	[**find that** S] #	from 4
6	S $\rightarrow$ [**find that** S] #	from 5, discharge 3
7	**Bel** (S $\rightarrow$ [**find that** S] #)	from 6
8	[**find that** S] **Bel** #	from 7, by rule about "find that S"

End Proof

A.4

$\models$ (**Bel** $\neg\exists$x. ϕ(x)) $\rightarrow$ [**find a** ϕ] **Bel** #

Proof:

1	**Bel** $\neg\exists$x. ϕ(x)	assumption
2	$\neg\exists$x. ϕ(x)	from 1
3	$\forall$x.$\neg\phi$(x)	from 2
4	$\neg\phi$(x)	from 3
5	ϕ(x)	assumption
6	#	from 4,5
7	[**find a** ϕ] (λy.#)(x)	from 6
8	ϕ(x) $\rightarrow$ [**find a** ϕ] (λy.#)(x)	discharge 5
9	$\forall$x.(ϕ(x) $\rightarrow$ [**find a** ϕ] (λy.#)(x))	from 8
10	**Bel** $\forall$x.(ϕ(x) $\rightarrow$ [**find a** ϕ] (λy.#)(x))	from 9
11	[**find a** ϕ] $\exists$x. **Bel** (λy.#)(x)	from 10, by rule about "**find a** ϕ"
12	[find a ϕ] **Bel** #	from 11

End Proof