Action Prediction Using a Mental-Level Model

Ronen I. Brafman
Stanford University
Dept. of Computer Science
Stanford, CA 94305-2140
brafman@cs.stanford.edu

Moshe Tennenholtz
Faculty of Industrial Engineering and Management
Technion
Haifa 32000, Israel
moshet@ie.technion.ac.il

Abstract

We propose a formal approach to the problem of prediction based on the following steps: First, a mental-level model is constructed based on the agent's previous actions; Consequently, the model is updated to account for any new observations by the agent, and finally, we predict the optimal action w.r.t. the agent's mental state as its next action. This paper formalizes this prediction process. In order to carry out this process we need to understand how a mental state can be ascribed to an agent, and how this mental state should be updated. In [Brafman and Tennenholtz, 1994b] we examined the first stage. Here we investigate a particular update operator, whose use requires making only weak modeling assumptions, and show that this operator has a number of desirable properties. Finally, we provide an algorithm for ascribing the agent's mental state under this operator.

Introduction

Tools for representing information about other agents are crucial in many contexts. Often, the goal of maintaining such information is to facilitate prediction of other agents' behavior, so that we can function better in their presence. Mental-level models, models that formalize the notion of a mental-state, provide tools for representing such information. Once we have a model of an agent's mental state, we can use it to predict future actions by finding out what an agent in such a state would perceive as its best action. The goal of this paper is to advance our understanding of basic questions related to the construction of a mental-level model, and in particular its application to prediction.

The idea of ascribing mental qualities for the purpose of prediction is not new. John McCarthy talks about it in [McCarthy, 1979]. One important aspect of McCarthy's ideas is that the agent's mental state is ascribed. That is, the entity being modeled may have in its structure nothing that resembles beliefs, desires, or other mental qualities, yet it may be possible and useful to model it *as if* it has such qualities. Thus, McCarthy views mental qualities as abstractions. This view is shared by another well-known author, Allen Newell[Newell, 1980], who contemplates the possibility of viewing computer programs at a level more abstract than that of the programming language, which he calls the *knowledge-level*.

The notion of a mental state is useful because it is abstract. Models at more specific levels, e.g., mechanical and biological models, are difficult to construct. They require information that we often do not have, such as the mechanical structure of the agent, or its program. On the other hand, mental-level models can be constructed based on observable facts – the agent's behavior – together with some background knowledge. In fact, as McCarthy points out, we might sometimes want to use these models even when we have precise lower level specifications of the agent, e.g. C code. This may be either because the mental-level description is more intuitive, or that computationally it is less complex to work with.

We present a formalism that attempts to make these ideas more concrete, and will hopefully lead to a better understanding of how the ascription of mental state could be mechanized. Motivated by work in decision-theory [Luce and Raiffa, 1957] and work on knowledge ascription[Halpern and Moses, 1990, Rosenschein, 1985], we suggested in [Brafman and Tennenholtz, 1994b] a specific structure for mental-level models, consisting of beliefs, desires and a decision criterion. This model showed how these elements act as constraints on the agent's action, and how these constraints can be used to ascribe beliefs to the agent *based* on its observed behavior. We would like to use this model in a particular prediction context, where we observe an agent performing part of a task, we know its goal, and we would like to predict its next actions. We use the following process: first, we ascribe beliefs to the agent based on the behavior we have seen so far. Next, we update the ascribed beliefs based on observations the agent makes, e.g., new information it has access to or the outcomes of its past actions. Then, to predict the agent's next action we examine what action would be perceived as best by an agent in this mental state.

In order to perform this prediction process we must understand how beliefs can be ascribed, how they should be updated, and how they should be used to determine the best perceived action. We have examined the first and the last question in[Brafman and Tennenholtz, 1994b] (although not in the context of prediction). In this paper we wish to concentrate on the second question, that of modeling the agent's belief change.

The reader should not confuse this last question with another important question which has received much attention: how should an agent change its beliefs given new information? (For example, see[Levesque, 1984, Friedman and Halpern, 1994, Katsuno and Mendelzon,

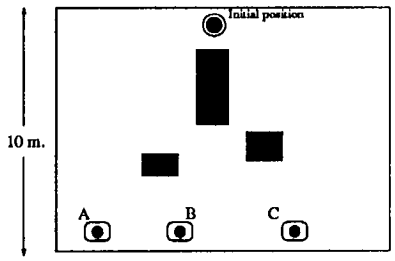

Figure 1: Example 1

1991, del Val and Shoham, 1993, Alchourron *et al.*, 1985, Goldzmidt and Pearl, 1992].) In our work we are concerned with externally modeling the changes occurring within the agent rather than saying how that agent should update its beliefs. Although that agent may be implementing one of the above belief revision methods, it is quite possible that it has no explicit representation of beliefs and that its "idea" of update is some complex assembler routine.

Our discussion of the problem of prediction will be in the context of the framework of mental-level modeling and belief ascription investigated in [Brafman and Tennenholtz, 1994b]. This framework is reviewed in Section 2. In Section 3 we discuss the problem of prediction. We suggest a three-step process for prediction and highlight the importance of the ascription of a belief change operator to this process. In Section 4 we introduce a particular belief change operator and show that it has desirable properties from a decision-theoretic perspective. Moreover, we show that under minimal assumptions, this belief change operator can always be ascribed to an agent. Finally, in Section 5 we discuss the algorithmic construction of beliefs obeying this belief change operator.

The Framework

We start by establishing a structure for mental-level models. Our framework, discussed in [Brafman and Tennenholtz, 1994b], is motivated by the work of Halpern and Moses[Halpern and Moses, 1990] and Rosenschein [Rosenschein, 1985] on knowledge ascription, and by ideas from decision-theory[Savage, 1972, Luce and Raiffa, 1957]. To clarify the concepts used we will refer to the following example.

Example 1 *We start with a robot located at an initial position. The robot is given a task of finding a small can located in one of three possible positions: A, B, or C. The robot can move in any direction and can recognize a can from a distance of 2 meters. (See Figure 1).*

The agent - basic description

An *agent* is described by a set of possible (local) states and a set of possible actions. The agent functions within an *environment*, which may also be in one of a number of states. We refer to the state of the system, i.e., that of both the agent and the environment as a *global state*. Without loss of generality, we will assume that the environment does not perform actions. The effects of the agent's actions are a (deterministic) function of its state and the environment's state.[1] This effect is described by the *transition function*. Together, the agent and the environment constitute a state machine with two components, with transitions at each state corresponding to the agent's possible actions. It may be the case that not all combinations of an agent's local state and an environment's state are possible, for example, if there is some correlation between the agent's and the environment's state. Those global states that are possible are called *possible worlds*.

Definition 1 *An* **agent** *is a pair* $\mathcal{A} = \langle L_{\mathcal{A}}, A_{\mathcal{A}} \rangle$, *where* $L_{\mathcal{A}}$ *is the agent's set of* **local states** *and* $A_{\mathcal{A}}$ *is its set of* **actions**. $L_{\mathcal{E}}$ *is the environment's set of possible states. A* **global state** *is a pair* $(l_{\mathcal{A}}, l_{\mathcal{E}}) \in L_{\mathcal{A}} \times L_{\mathcal{E}}$. *The set of* **possible worlds**, S, *is a subset of the set of global states* $L_{\mathcal{A}} \times L_{\mathcal{E}}$. *A* **context**[2] $C = \langle \tau, I \rangle$, *consists of a* **transition function**, $\tau : (L_{\mathcal{A}} \times L_{\mathcal{E}}) \times A_{\mathcal{A}} \rightarrow (L_{\mathcal{A}} \times L_{\mathcal{E}})$, *and the set* $I \subset W$ *of possible initial states.*

Example 1 *(continued): Suppose that our robot has imperfect sensing of its position. Thus, its local state would include its position reading as well as whether or not it has observed the can; its actions correspond to motions in various directions. The state of the environment describes the* actual *position of the can and each possible world describes (1) the robot's position, (2)the can's position (3) the robot's position reading and (4) whether it has observed the can. The transition function describes how each motion changes the global state of the system. There are three initial states. In each the position of the robot is the given initial position and the can is located in one of positions A, B, or C.*

We say that an agent *knows* some fact if in all the worlds the agent should consider possible, this fact holds. The worlds an agent should consider possible are those in which its information (as represented by its local state) would be as it is now. An agent can distinguish between two worlds in S if and only if *its* state in them is different. Therefore, an agent whose local state is l can rule out as impossible all worlds in which his local state would have been different, but cannot rule out worlds in S in which his local state would have been l. Thus, knowledge corresponds to what holds in all worlds that the agent cannot distinguish from the actual world.

Definition 2 *The set of* **worlds possible at** l, $PW(l)$, *is* $\{w \in S$: *the agent's local state in* w *is* $l\}$. *The agent* **knows** φ *at* $w \in S$ *if* φ *holds in all worlds in* $PW(l)$, *where* l *is its local state at* w.

Example 1 *(continued): Let us assume from now on that the robot's position reading is perfect. In that case, the robot knows its position, since a position reading r is*

[1] A framework in which the results of actions are nondeterministic and in which the environment may take actions as well can be mapped into this framework using richer state descriptions and larger sets of states, a common practice in game theory.

[2] Though context is an overloaded term, its use here seems appropriate, following [Fagin *et al.*, 1994].

part of its local state and r can only be obtained in worlds in which the actual position of the robot is r. However, unless the robot has observed the can, it does not know the can's position, since it has possible worlds in which the can's position is different.

The agent's observed, or programmed behavior is described by a protocol, which is a function that assigns an action to each state.

Definition 3 *A* **protocol** *for an agent $\mathcal{A}$ is a function $\mathcal{P}_{\mathcal{A}} : L_{\mathcal{A}} \rightarrow A_{\mathcal{A}}$.*

Example 1 *(continued): The robot's protocol would specify in what direction to head in each position.*

Later, we will consider the problem of *prediction* in the following context: We have observed an agent taking a number of actions in pursuit of some known goal and we would like to predict its next action. In this setting, the behavior of the agent involves taking a number of steps aimed at achieving some goal. To model this we need some formal notion that will describe the dynamic evolution of the agent in its environment. We call this a *run*.[3] A run is a sequence of global states that commence at an initial state, such that each following state can be achieved by executing some action at the previous state. We also have special notations for more specific types of partial runs, in which the kind of actions allowed are restricted by some given protocol. A $(w, \mathcal{P})$ run prefixes/suffixes are simply prefixes/suffixes of runs that end/begin in a state w, throughout which the agent acts in accordance with the protocol $\mathcal{P}$.

Definition 4 *A* **run** *of an agent $\mathcal{A}$ whose possible initial worlds are I is a sequence of possible worlds, $r = \{w_0, w_1, \ldots, w_n\}$ satisfying the following conditions: (1) $w_0 \in I$; and (2) $w_{i+1} = \tau(w_i, a)$, where $a \in A_{\mathcal{A}}$; The set of all possible runs is denoted by $\mathcal{R}$. The runs possible at l, $PR(l)$, are those runs in $\mathcal{R}$ in which at some stage the agent's local state is l.*

A $(w, \mathcal{P})$ **run-prefix** *of an agent $\mathcal{A}$ is a run prefix $\{w_0, w_1, \ldots, w_m = w\}$ such that $w_{i+1} = \tau(w_i, \mathcal{P}(w_i))$ for $0 \leq i < m$. A $(w, \mathcal{P})$* **run-suffix** *of an agent $\mathcal{A}$ is a run suffix $\{w_k = w, \ldots, w_l\}$, such that $w_{i+1} = \tau(w_i, \mathcal{P}(w_i))$ for $k \leq i < l$.*

Example 1 *(continued): A run is a sequence of global states. In our example this sequence can be described by the trajectory of the robot through the space, the position of the can, and, at each point along the run, whether the robot has observed the can.*[4] *Each such trajectory must start at the initial position.*

The Agency Hypothesis

What we described until now provides a basic layer on top of which we construct the ascribed mental state of the agent. This state will relate three key notions: beliefs, goals and decision criteria. This state will manifest itself in the agent's behavior.

[3] We will assume runs are finite. The extension to infinite runs is straightforward.

[4] A continuous model of time may be preferred here. This is possible, e.g., [Brafman *et al.*, 1994].

We start the description of the mental level model by defining the notion of belief. Belief is part of an abstract description of the agent's state. It sums up the agent's view of the world, and is a basis for decision making. Therefore, we make belief a function of the agent's *local* state, represented by a *belief assignment*, which assigns to each local state a nonempty subset of the set of possible worlds, the worlds the agent considers *plausible*.

Definition 5 *A* **belief assignment** *is a function, $B : L_{\mathcal{A}} \rightarrow 2^S$, such that for all l : $B(l) \neq \emptyset$ and if $w \in B(l)$ then (1) the agent's local state in w is l, and (2)there exists a $(w, \mathcal{P}_{\mathcal{A}})$ run-prefix.*

As we see, the worlds the agent considers plausible must be consistent with the agent's past actions.

Example 1 *(continued): At each local state in which the can has not been observed yet, the robot has three possible worlds. Each corresponds to a different position of the can. A belief assignment would assign a subset of these at each local state. If $B(l)$ contains the world in which the can is at A, the robot is viewed as believing that the can is in location A.*

Knowledge (or $PW(l)$) defines what is theoretically possible; belief defines the set of worlds that, from the agent's perspective, should be taken into consideration.This notion of belief makes sense only as part of a fuller description of the agent's mental level. Such a description requires additional notions, which we now introduce. We start with the agent's preference order over the set of run suffixes, represented by a *utility function*. This preference order embodies the relative desirability of different futures.

Definition 6 *A* **utility** *function u is a real-valued function on the set of run-suffixes.*

It is well known [von Neumann and Morgenstern, 1944] that a utility function can represent preference orders satisfying certain assumptions, which in this paper we will accept. This means that for any two run suffixes r_1, r_2: r_1 is preferred over r_2 *iff* $u(r_1) > u(r_2)$. We would also expect additional properties from u. These properties would capture our intuitions that certain related suffixes should have similar utility. These consideration are tangent to our current discussion.

Example 1 *(continued): In our example, we will assume a simple arbitrary utility function that depends on the length of the robot's trajectory and the location of the can. If the length of the robot's trajectory is x, then*
$u = 10 - x + 20*$(The trajectory terminates at the can)

When the exact state of the world is known, the result of following some protocol, $\mathcal{P}$, is also precisely known. (Actions have deterministic effects). We can then evaluate a protocol by looking at the utility of the run it would generate in the actual world. However, due to uncertainty about the state of the world, the agent considers a number of states to be possible. It can then subjectively assess $\mathcal{P}$ in a local state l by a vector whose elements are the utilities of the plausible runs $\mathcal{P}$ gener-

ates. More precisely we have the following definition, in which we assume the set $B(l)$ is ordered.[5]

Definition 7 *Given a context C and a belief assignment B, let w_k denote the k^{th} state of $B(l)$. The* **perceived outcome** *of a protocol $\mathcal{P}$ in l is a tuple whose k^{th} element is the utility of the $(w_k, \mathcal{P})$ run-suffix.*

Example 1 *(continued): Suppose that the possible worlds of our example are ordered alphabetically according to the position of the can. Suppose that $B(l_I) = \{A, B\}$, i.e., initially, the robot believes the can is either in A or in B. The perceived outcome of the protocol that takes the robot to A first, and if not there to B, is $\{19, 15\}$, since the distance to A is (approx.) 11 meters and the distance from A to B is (approx.) 4 (so the total distance is 15). Notice that the perceived outcome disregards possible worlds that are not plausible.*

While utilities are easily compared, it is not a-priori clear how to compare perceived outcomes, thus, how to choose among protocols. A strategy for choice under uncertainty is required. This strategy could depend on, for example, the agent's attitude towards risk. This strategy is represented by the *decision criterion*, a function taking a set of perceived outcomes and returning the *set* of most preferred among them.

Definition 8 *A* **decision criterion** *ρ is a function which maps each set of tuples (of equal length) of real numbers, to a subset of it.*

Two examples of decision criteria are *maximin*, which chooses the tuples in which the worst case outcome is maximal, and the *principle of indifference* which prefers tuples whose average outcome is maximal.(A fuller discussion of decision criteria appears in [Luce and Raiffa, 1957, Brafman and Tennenholtz, 1994a].)

We come to a key definition that ties all of the components we have discussed so far.

Definition 9 *The* **agency hypothesis***: at each state the agent follows a protocol whose perceived outcome is most preferred (according to its decision criterion) among the set of perceived outcomes of all possible protocols.*[6]

The agency hypothesis takes the view of a rational balance among the agent's beliefs, utilities, decision criterion and behavior. It states that the agent chooses actions whose perceived outcome is maximal according to its decision criterion.

Since given a fixed utility function the decision criterion induces a choice among actions, we will often use the term 'most preferred protocol' in place of 'the protocol whose perceived outcome is most preferred'.

Ascribing Belief

The various elements at the mental-level are related through a rational balance. We can exploit this relation to ascribe a mental state to an agent. We use the available information, such as observed behavior and background information, to constrain the possible values of the unknown mental state. We now show how belief can be ascribed according to our framework.

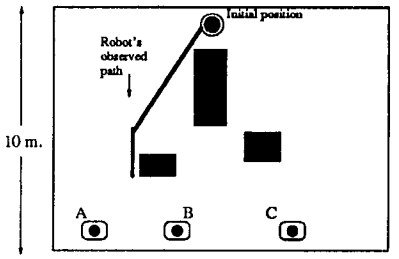

Figure 2: Robot's initial path

Belief ascription requires certain information regarding the agent. This information should specify some of the other elements of the rational balance we have just discussed. Using this information we look for belief assignments confirming the agency hypothesis. That is, suppose the entity modeled satisfies the agency hypothesis and that its utilities and decision criterion are as given, then such beliefs would lead us to act as was observed. This is a process of constraint satisfaction. Thus a formal statement of belief ascription is the following:

> Given a context C for an agent $\mathcal{A}$, find a belief assignment B such that B together with the agent's behavior, its utility function and its decision criterion confirm the agency hypothesis.

Example 1 *(continued): We will try to ascribe beliefs to the robot. We assume that the decision criterion used is maximin. (What follows applies also to the principle of indifference). Suppose we observe the robot moving along the path described in Figure 2. What can we say about its beliefs? We must see under what beliefs the path observed would yield the highest utility. It is easy to see the ascribed plausible worlds are $\{A, B\}$. If the robot believed only one state to be plausible it would head directly to it. Similarly, if $\{B, C\}$ was believed the robot's path would head more toward them. If the robot believes $\{A, B, C\}$, a better path would be along the middle, rather than the left-hand side.*

Our ability to ascribe belief in the framework of the mental-level model just presented is thoroughly discussed in [Brafman and Tennenholtz, 1994b].

Predictions

We wish to explore the application of mental-level models in a particular form of prediction: We observed an agent taking part in some activity; we know its goals; and we wish to predict its next actions. In what follows we try to examine what problems this task raises and how we might solve them. We will concentrate on one particular issue, belief change. We will soon see how it relates to our task. The approach we suggest underlies some of the work done in belief and plan ascription. We believe that a formal approach will aid in understanding

[5] This is used to simplify presentation. All definition extends to infinite sets by replacing tuples with functions.

[6] The possible protocols are implicitly defined by the set of actions $A_{\mathcal{A}}$ (cf. Def. 1).

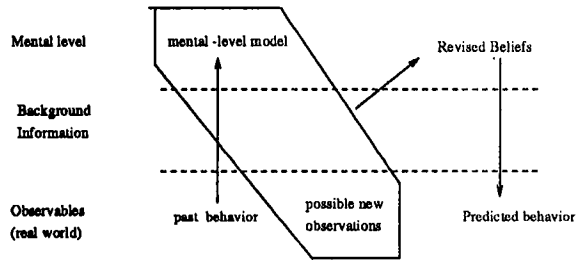

Figure 3: Three step prediction process

this task better and in detecting the implicit assumptions made in predicting an agent's future behavior.

To predict an agent's next action we go through three steps (illustrated in Figure 3): (1) construct a mental-level model of the agent based on actions performed until now; (2) revise the agent's ascribed beliefs, if needed, based on the observations it made after performing the last action; (3) predict the action which has the most preferred perceived outcome based on these beliefs.

Example 1 *(continued): In the previous section we saw an example of belief ascription. This corresponds to the first stage: constructing a mental-level model based on observations and background knowledge. The robot's beliefs were* $\{A, B\}$. *Based on this we can predict that the robot would continue to move in its current direction until it can observe whether the can is in A or B. Suppose the can was observed to be in B. In that case, the beliefs of the robot are revised to contain only B. Given these beliefs, we expect the robot to turn to the right (i.e. toward B).*

Our human experience shows that models of mental-state are useful in predicting human behavior, and we believe they are also likely to succeed with human-made devices (hence the *agency* hypothesis: the device acts as an agent of its designer, echoing its goals and beliefs). Thus, using mental-level models seems to make heuristic sense. However, when is this really appropriate? Moreover, when is the particular formalism suggested here appropriate? Reexamining the three-step prediction process we see two major implicit assumptions:

- We can model the observed behavior of an agent using a mental-level model.
- We can assume some methodical belief change process.

We discussed the first among these issues in our previous work [Brafman and Tennenholtz, 1994b]. In particular, we have shown a class of agents that can be ascribed the mental-level model discussed in Section 2. We devote the rest of this paper to the second issue.

Belief change

Suppose we have constructed a mental-level model based on past behaviors. To use it in predicting future behavior, we must make an additional assumption, that there is some temporal coherence of beliefs. Consider the example of the robot that accompanied the preceding sections. We observe the robot move along a certain path and ascribe it the belief assignment $\{A, B\}$. At a certain stage it is near enough to A and B to be able to see whether the can is in one of these two positions. We expect this new information to affect the behavior of the robot. In our ascribed model of the robot, we expect this information to be manifested in terms of belief change. However, unless the new belief can be somehow constructed from the old beliefs and the observation, we will have very little ability to predict future behavior.

We first suggest a restriction on the relationship between beliefs in different states. Later on we will show that this restriction is both natural and useful.

Admissibility

Consider the following restriction: if my new information does not preclude all of the runs I previously considered plausible, I will consider plausible all runs previously considered plausible that are consistent with my new information.

Let $N_{\mathcal{A}}(s) \stackrel{\text{def}}{=} \tau(\mathcal{P}_{\mathcal{A}}(s))$, i.e., the state that will follow s when $\mathcal{A}$ performs the action specified by its protocol, and $N_{\mathcal{A}}(T) = \{N_{\mathcal{A}}(s) | s \in T\}$.

Definition 10 *A belief assignment B (for agent* $\mathcal{A}$*) is* **admissible**, *if for local states* l, l' *such that* l' *follows* l *on some run: whenever* $N_{\mathcal{A}}(B(l)) \cap PW(l') \neq \emptyset$ *then* $B(l') = N_{\mathcal{A}}(B(l)) \cap PW(l')$; *otherwise* l' *is called a* **revision state** *and* $B(l')$ *can be any subset of* $PW(l')$.

If we were to assume that the worlds here are models of some theory then, in syntactic terms, admissibility corresponds to conjoining the new data with the existing beliefs, whenever this is consistent. It is closely related to the probabilistic idea of conditioning beliefs upon new information.

It turns out that admissible belief assignment can be viewed in a different way. As the following theorem shows an admissible belief assignment is equivalent to a belief assignment induced by a ranking of the set of initial states, that is, a belief assignment which assigns to every local state those worlds in $PW(l)$ that originate in initial states whose rank is minimal. Intuitively, we associate minimal rank with greater likelihood.

Theorem 1 [7] *Assuming perfect recall,*[8] *let* $I_{(w,\mathcal{P})}$ *denote the initial state of the* $(w, \mathcal{P})$ *run prefix. A belief assignment B is admissible iff there is a ranking function r (i.e., a total pre-order) on the possible initial worlds I, such that* $B(l) = \{w \in PW(l) : I_{(w,\mathcal{P})}$ is r-minimal$\}$.

Why admissibility

The fact that admissible beliefs have a nice representation seems encouraging. It suggests a refinement to our model in which beliefs have the additional structure provided by a ranking over possible worlds. However, this by itself is no reason to accept this restriction. Remember that we want to show that the mental-level model is an abstraction that is grounded in lower level phenomena. The kind of support we need would look like "under

[7] Proofs are omitted due to lack of space, and will appear in a longer version of this paper.

[8] An agent is said to have perfect recall if its local state contains all previous local states.

assumption X on the agent's behavior, a ranked belief assignment can be ascribed to it". This is what we wish to do in this section. Once these questions are answered, we would be able to make justified predictions based on the approach presented in the previous section. On our way to this goal, we will also get some interesting results from a decision-theoretic perspective.

Recall the agency hypothesis. The agent was viewed as choosing among protocols based on the utility of the runs they generate[9] and its beliefs. However, there is an alternative way for choosing among actions given the agent's beliefs, which is popular in decision-theory.

Definition 11 *A mental-level model of an agent is said to have the* **backwards-induction** *property if the following holds: for every local state l, the observed protocol $\mathcal{P}$ is at least as preferred as any of the protocols which differ from $\mathcal{P}$ only on l. We call $\mathcal{P}$ the backwards-induction protocol.*

This is called the backwards induction property because we can construct a protocol $\hat{\mathcal{P}}$ that satisfies it using backwards induction. If all the children of l are final states, then let $\hat{\mathcal{P}}(l)$ be the most preferred action at l. (In this case each action determines a run suffix.) Having specified what to do at the next-to-last local states, we now assign their parents an action that will give a most preferred run suffix, given our existing choice for the children. We repeat this process until we reach the initial states I.

Lemma 1 *$\hat{\mathcal{P}}$ are the only protocols satisfying the backwards-induction property.*

Backwards induction gives a very thorough manner of choosing a protocol, and it is perhaps the most popular way to determine rational action in a decision-theoretic setting. This point will become useful later.

Another decision-theoretic concept we will use is the following (where $\circ$ denotes vector concatenation):

Definition 12 *A decision criterion satisfies the* **sure-thing** *principle if $v \circ v'$ is at least as preferred as $u \circ u'$ whenever v is at least as preferred as v' and u is at least as preferred as u'.*

That is, suppose the agent has to choose between two actions, a and a'. It prefers a over a' when the plausible worlds are B. It also prefers a over a' when the plausible worlds are B'. If this agent satisfies the sure-thing principle it should also prefer a over a' when the plausible worlds are $B \cup B'$.

In what follows we assume that the agent *satisfies* the sure-thing principle. Given the above machinery, first we look at normative reasons for accepting admissibility.

Theorem 2 *Let $\mathcal{A}$ be an agent with admissible beliefs. If $\mathcal{P}$ is its most preferred protocol at the initial local state l_i, then $\mathcal{P}$ will still be most preferred at all the following states.*

[9]This section assumes that there are only a finite number of possible local states, that runs are finite, and that the agent has perfect recall. We will also use the term "most preferred" loosely, meaning "one of the most preferred".

This means that the agents of Theorem 2 choose a protocol once and for all in l_i based on the protocol's perceived outcome. When beliefs are not admissible, this need not necessarily be the case. (A counter example is easy to construct.)

Another nice property associated with admissible beliefs is given by the following theorem and corollary.

Theorem 3 *Let $\mathcal{A}$ be an agent with admissible beliefs, a protocol that has backwards-induction property for $\mathcal{A}$ is also most preferred by $\mathcal{A}$ at the initial state.*

Corollary 1 *Let $\mathcal{P}_{BI}$ satisfy the backwards induction property. There is a utility function on* states *such that an agent with admissible beliefs executing the best local action at each state, will perform $\mathcal{P}_{BI}$. Consequently, this protocol is most preferred at the initial state.*

We get an interesting property of agents whose actions are consistent with a backwards induction protocol and who satisfy the sure-thing principle. These agents can be *viewed* as if they act based on admissible beliefs, even if their beliefs are actually not admissible. This boosts the credibility of the assumption that belief change is admissible because it shows a basic and important class of agents that do not use admissible beliefs, yet, we can legitimately *model* them as having admissible beliefs.

Our claim that admissibility is a good modeling assumption is strengthened by the following result:

Theorem 4 *Let $\mathcal{P}$ be the observed protocol of an entity. If this entity can be ascribed beliefs at the initial state and at subsequent revision states based on this protocol, it can be ascribed an admissible belief assignment at all local states.*

Thus, admissibility is almost free as long as we can ascribe beliefs at the initial state, based on the observed protocol.

The previous results imply that admissible beliefs are useful for ascription and prediction. In fact, the results can be even further improved. The fact that we associate utilities with runs rather than states complicates our life when we try to ascribe beliefs. To ascribe beliefs we must at each state compare whole protocols and the run suffixes they produce. It would be much easier if we could only look at single actions and their immediate outcomes. This would require defining a utility function over the set of states, rather than the set of run suffixes. Indeed, this is possible:

Theorem 5 *Let $\mathcal{P}$ be the observed protocol of an agent, and suppose that this agent can be ascribed beliefs at the initial state and at all subsequent revision states based on this protocol. Then, it can be ascribed an admissible belief assignment at all local states and a local utility function* over states *such that its observed* action *has the most preferred perceived outcome according to the local utility function.*

Constructive Ascription

In this section we offer a more constructive account of admissible belief ascription. Any ascription algorithm will have to come up with a belief assignment that satisfies the agency hypothesis at a given local state. That

is, it would have to generate for a given local state beliefs that would have made the agent act as it did. We assume such an algorithm as a subroutine (see [Brafman and Tennenholtz, 1994a]). What we offer here is an algorithm for generating an admissible belief assignment given this subroutine. We show that the complexity of obtaining global consistency is a low polynomial in the complexity of computing a locally consistent assignment in a given local state. The algorithm works under the assumption that the agent obeys the sure-thing principle and that it has perfect recall.

An added feature of the algorithm is that it returns a special belief assignment, which we call the *most general belief assignment* (mgb). We discuss the mgb in [Brafman and Tennenholtz, 1994b]. Intuitively, it is the belief assignment that makes fewest assumptions. In terms of the ranking that it ascribes to the agent, it makes worlds as minimal as possible.

Theorem 6 *Let c be the cost of computing a most general belief assignment in a local state, let i be the number of possible initial (global) states and let m be the number of local states that can be encountered during the execution of the protocol. It is possible to construct an mgb in time $O(i^2 \cdot m \cdot c)$.*

Proof (sketch): Let l^i be an initial local state. The procedure is as follows:

1. Set $b = PW(l^i)$
2. If $b = \emptyset$ then stop. There is no admissible belief assignment.
3. Set b to the maximal locally consistent belief assignment (w.r.t. to l^i) that is a subset of b.
4. Start constructing the tree of local states consistent with b (i.e., such that their possible worlds contain worlds in b). For each such l assign $B(l) = PW(l) \cap b$.
5. If for any state l the assignment $B(l)$ is locally inconsistent with the decision criteria, let b' equal the maximal consistent local assignment for that state that is contained in $B(l)$. Let $b := b \setminus (B(l) \setminus b')$. Goto 2.
6. Otherwise, let $B(l^i) = b$. Delete all the states for which beliefs were assigned. You may be left with one or more trees. Repeat the algorithm for each of them.

It is easy to see that the algorithm is $O(i^2 \cdot m \cdot c)$. The most we repeat step 3 for any revision state is i and there can be no more than i revision states. Since step 4 performs a calculation of time $O(m \cdot c)$, the total time is $O(i^2 \cdot m \cdot c)$. In addition, we can show:

Lemma 2 *The above algorithm returns the mgb, if one exists.*

Discussion

A question that motivates much of the research in belief and belief change [Levesque, 1984, Friedman and Halpern, 1994, Katsuno and Mendelzon, 1991, del Val and Shoham, 1993, Alchourron *et al.*, 1985, Goldzmidt and Pearl, 1992] is the following: Given that we can make better programs by equipping them with large amounts of knowledge, how should this knowledge be represented, and how should it be updated? This work often relies on our intuitive notion of belief and belief change. Often it is implicitly assumed that it is we, the designers, that will supply the agent with its knowledge, at least initially.

We are concerned with a more specific question of representation and ask: how should an agent represent its information about *another* agent in a way that will facilitate explaining and predicting the other agent's behavior. Moreover, we assume that the bulk of an agent's knowledge about other agents comes from a particular source: observation of these agent's behavior. Thus, we are more concerned with modelling agent's ascribed beliefs, than with designing them.

An important related work that shares some of our perspective is Levesque's work [Levesque, 1986], which is concerned with treating computers as believers. However, this work describes the beliefs of one particular class of agents whose actions are answering queries. Our work attempts to address a more general class of agents, whose actions are arbitrary.

Modelling data is a central task of machine-learning. Much like our work, these models are constructed to help make predictions, e.g., a decision tree helps us predict what class an instance belongs to, or the detection algorithm of [Cohen *et al.*, 1995]. Our work brings to this task a special bias in the form of the agency-hypothesis: Machines are agents of their designers; they are usually designed with a purpose in mind and with some underlying assumptions; therefore, they should be modeled accordingly. With this motivation in mind, this work and [Brafman and Tennenholtz, 1994b] attempt to understand the basis for modeling entities *as if* they have a mental state. The central issues are: what elements should such a model contain? How should we use observable information to construct it? And, under what assumptions is our modeling "bias" justified?

This paper complements our previous work on belief ascription [Brafman and Tennenholtz, 1994b] and helps supply initial answers to the above-mentioned questions. In this paper, we reviewed our proposal for the structure of mental-level models and their construction, and explained how they can be used to predict an agent's future behavior. In order to use mental-level models in making predictions, we must constantly update them. A key component in this update process is the ability to *model* the belief change of other agents. We suggested admissibility as a belief change operator, examined its properties and showed that we can accept it under rather weak modeling assumptions. Putting these ingredients together, we get a theory of action prediction using a mental-level model, which consists of the three-step process, a theory of belief ascription (discussed in [Brafman and Tennenholtz, 1994b]), and a study of belief change modelling.

Acknowledgement

We are grateful to Yoav Shoham and Nir Friedman for their valuable comments on this work.

Ronen Brafman was supported in part by AFOSR grant AF F49620-94-1-0090.

References

[Alchourron *et al.*, 1985] Alchourron, C. E.; Gärdenfors, P.; and Makinson, D. 1985. On the logic of theory change: partial meet functions for contraction and revision. *Journal of Symbolic Logic* 50:510–530.

[Brafman and Tennenholtz, 1994a] Brafman, R. I. and Tennenholtz, M. 1994a. Belief ascription. Technical report, Stanford University.

[Brafman and Tennenholtz, 1994b] Brafman, R. I. and Tennenholtz, M. 1994b. Belief ascription and mental-level modelling. In Doyle, J.; Sandewall, E.; and Torasso, P., editors 1994b, *Proc. of Fourth Intl. Conf. on Principles of Knowledge Representation and Reasoning.* 87–98.

[Brafman *et al.*, 1994] Brafman, R. I.; Latombe, J. C.; Moses, Y.; and Shoham, Y. 1994. Knowledge as a tool in motion planning under uncertainty. In Fagin, R., editor 1994, *Proc. 5th Conf. on Theor. Asp. of Reas. about Know.*, San Francisco. Morgan Kaufmann. 208–224.

[Cohen *et al.*, 1995] Cohen, P.; Atkin, M.; Oates, T.; and Gregory, D. 1995. A representation and learning mechanism for mental states. In *Proc. of the AAAI Spring Symposium on Representing Mental States and Mechanisms.*

[del Val and Shoham, 1993] del Val, A. and Shoham, Y. 1993. Deriving properties of belief update from theories of action. In *Proc. Eleventh Intl. Joint Conf. on Artificial Intelligence (IJCAI '89).* 584–589.

[Fagin *et al.*, 1994] Fagin, R.; Halpern, J. Y.; Moses, Y.; and Vardi, M. Y. 1994. *Reasoning about Knowledge.* MIT Press. to appear.

[Friedman and Halpern, 1994] Friedman, N. and Halpern, J. Y. 1994. A knowledge-based framework for belief change. Part I: Foundations. In *Proc. of the Fifth Conf. on Theoretical Aspects of Reasoning About Knowledge*, San Francisco, California. Morgan Kaufmann.

[Goldzmidt and Pearl, 1992] Goldzmidt, M. and Pearl, J. 1992. Rank-based systems: A simple approach to belief revision, belief update and reasoning about evidence and actions. In *Principles of Knowledge Representation and Reasoning: Proc. Third Intl. Conf. (KR '92).* 661–672.

[Halpern and Moses, 1990] Halpern, J. Y. and Moses, Y. 1990. Knowledge and common knowledge in a distributed environment. *J. ACM* 37(3):549–587.

[Katsuno and Mendelzon, 1991] Katsuno, H. and Mendelzon, A. 1991. On the difference between updating a knowledge base and revising it. In *Principles of Knowledge Representation and Reasoning: Proc. Second Intl. Conf. (KR '91).* 387–394.

[Levesque, 1984] Levesque, H. J. 1984. A logic of implicit and explicit belief. In *Proc. National Conf. on Artificial Intelligence (AAAI '84).* 198–202.

[Levesque, 1986] Levesque, H. J. 1986. Making believers out of computers. *Artificial Intelligence* 30:81–108.

[Luce and Raiffa, 1957] Luce, R. D and Raiffa, H. 1957. *Games and Decisions.* John Wiley & Sons, New York.

[McCarthy, 1979] McCarthy, J. 1979. Ascribing mental qualities to machines. In Ringle, M., editor 1979, *Philosophical Perspectives in Artificial Intelligence*, Atlantic Highlands, NJ. Humanities Press.

[Newell, 1980] Newell, A. 1980. The knowledge level. *AI Magazine* 1–20.

[Rosenschein, 1985] Rosenschein, S. J. 1985. Formal theories of knowledge in AI and robotics. *New Generation Comp.* 3:345–357.

[Savage, 1972] Savage, L. J. 1972. *The Foundations of Statistics.* Dover Publications, New York.

[von Neumann and Morgenstern, 1944] Neumann, J.von and Morgenstern, O. 1944. *Theory of Games and Economic Behavior.* Princeton University Press, Princeton.