Towards Action Prediction Using a Mental-Level Model

Ronen I. Brafman
Dept. of Computer Science
Stanford University
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; next, 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 and show that its ascription requires making only weak modeling assumptions.

1 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 use formal counterparts of various mental states to describe the state of an agent, 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 discusses it in [McCarthy, 1979]. An important aspect of his approach is that even when nothing in the internal structure of the entity modeled directly resembles beliefs, desires, or other mental qualities, 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. We might do this either because the mental-level description is more intuitive or because computationally it is less complex to work with.

We present a formalism that attempts to make these ideas more concrete and that will hopefully lead to 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. 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, in order 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.

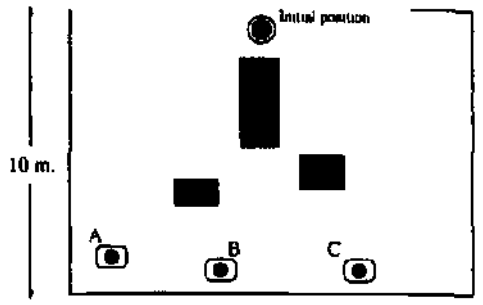

Figure 1: Example 1

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, 1991; del Val and Shoham, 1993; Alchourron *et a*/., 1985; Goldszmidt 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.

2 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).*

2.1 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.* W.l.o.g., 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. Those global states that are possible are called *possible worlds.*

Definition 1 *An* **agent** *is a pair* $\mathcal{A} = (L_{\mathcal{A}}, A_{\mathcal{A}})$, *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* $I \subset S$ *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, (S)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.

Definition 2 *The set of* worlds possible at *l*, *PW(l), is {w € S : the agent's local state in w is l}. The agent* knows *Φ at w € 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 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 formally captured by the notion of a protocol.

[1] A framework in which the environment does act 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].

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 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]

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 $k \leq i < l$.*

A $(w, \mathcal{P})$ run prefix is a prefix of a run starting at I and ending at w, throughout which the agent acts in accordance with the protocol $\mathcal{P}$. $(w, \mathcal{P})$ run suffixes are defined analogously, where w is now the initial state of the run suffix.

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.*

2.2 The Agency Hypothesis

We start the description of mental-level models by defining the notion of a belief assignment.

Definition 5 *A* **belief assignment** *is a function, $B : L_{\mathcal{A}} \rightarrow 2^{\mathcal{S}}$, 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 $(w, \mathcal{P}_A)$ run-prefix.*

A belief assignment specifies those worlds the agent considers currently plausible. These worlds should be consistent with the agent's past actions. The notion of plausible runs at l is similar.

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.[5] 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 $\mathcal{P}$ generates. More precisely we have the following definition, in which we assume the set $B(l)$ is ordered.[6]

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 ignores possible worlds that are not plausible.*

[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].

[5] We remark, that (after adding interpretations to each world) this approach yields a $KD45$ belief operator.

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

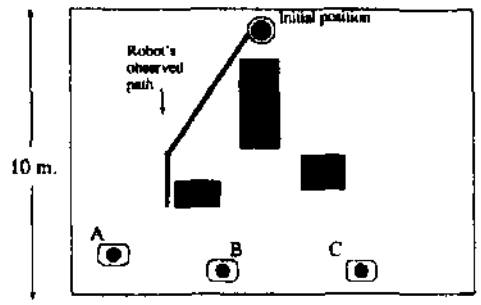

Figure 2: Robot's initial path

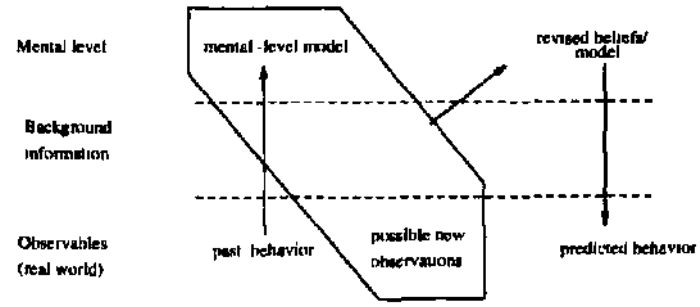

Figure 3: Three step prediction process

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 *p 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.*[7]

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'.

2.3 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.

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 *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 discussed in [Brafman and Tennenholtz, 1994b].

3 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 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.

[7] The possible protocols are implicitly denned by the set of actions A_A (cf. Def. 1).

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 these beliefs we can predict that the robot will 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.

4 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.

4.1 Admissibility

Consider the following restriction: if my new information is consistent with some of the runs I previously considered plausible, I will now consider plausible those runs previously considered plausible that are consistent with this new information.

Let *NA{S) = T(VA(S)),* i.e., the state that will follow *s* when *A* performs the action specified by its protocol, *and* $N_A\{T) = \{N_A(s)\backslash s \in T\}$.

Definition 10 *A belief assignment B (for agent A) is* admissible, *if for local states 1,1' such that I' follows I on some run: whenever NA(B(1))* n *PW(I')* 0 *then B{V) =* $N_A(B(l))$ D *PW{(li)j otherwise V is called a* revision state *and B(V) can be any subset of PW(V).*

If worlds corresponded to 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 plausibility.

Theorem 1 [8] *Assuming perfect recall,*[9] *let I. -p, denote the initial state of the (w,V) 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 € PW{1) : I, (w,p).* is r-minimal}.

4.2 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 mental-level models are abstractions that are grounded in lower level phenomena. The kind of support we need would look like "under assumption *X* on the agent's behavior, a ranked belief assignment can be ascribed to it". In this section, we would like to present results of this nature. 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[10] and its beliefs. However, there is an alternative way for choosing among actions given the agent's beliefs, called *backwards induction.*

Definition 11 *A* backwards induction *(BI) protocol for an agent A is defined inductively as follows: For local states I, all of whose children are final states, assign a most preferred action at that state. (In this case an action determines a run suffix.) Inductively, assign to each local state an action that is most preferred given the choices for its descendents.*

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

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

[10] 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.

Backwards induction is considered the rational way of choosing actions according to classical decision-theory. Another decision-theoretic concept we will use is the following (where o denotes vector concatenation):

Definition 12 *A decision criterion satisfies the* sure-thing *principle if v o v′ is at least as preferred as u* o *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* U *B′*. In what follows we assume that the agent *satisfies* the sure-thing principle.

One final note: when we compare protocols at an initial state we only care about their outcome on states that are plausible; we are indifferent to what actions we take in revision states. We can view this as a choice among partial protocols, defined only on the plausible worlds. However, once we get to a revision state we will have to choose among a new set of partial protocols, depending upon our beliefs in the revision state. Since we assume perfect recall, these choices are independent. Hence, it will be convenient for us to think about the agent making all these choices initially. That is, at the initial state it chooses not only what to do on the plausible worlds, but also what to do on revision states, if it ever gets to them. This way we view the agent as choosing among full protocols, and the notion of most preferred protocol will be defined accordingly.

Given the above machinery, we first look at normative reasons for accepting admissibility.

Theorem 2 *Let A be an agent with admissible beliefs. Its most preferred protocols at the initial local state remain most preferred at all the following states.*

Therefore, agents of Theorem 2 can choose a protocol once and for all at the initial state based on its perceived outcome. When beliefs are not admissible, a counter example can be constructed where protocols most preferred at the initial states are not most preferred later on.

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

Theorem 3 *Let A have admissible beliefs, V is a BI protocol for A iff it is most preferred at the initial state.*

Corollary 1 *Assume A has admissible beliefs. There is a utility function on* states *such that A can be viewed as executing the best local action at each state.*

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

Theorem 4 *Let V 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 free if we can ascribe beliefs at the initial and revision states.

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. Doing 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 *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.*

5 Discussion

The following question motivates much of the research in belief and belief change: 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? (For example, see [Levesque, 1984; Friedman and Halpern, 1994; Katsuno and Mendelzon, 1991; del Val and Shoham, 1993; Alchourron *et al,* 1985; Goldszmidt and Pearl, 1992].) That work often attempts to capture our intuitive notion of belief and belief change. In addition, it often implicitly assumes that we, the designers, are those who 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 modeling agent's ascribed beliefs than with designing them.

An important related work that shares some of our perspective is Levesque's [Levesque, 1986], which is concerned with treating computers as believers. However, his 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.

Modeling 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. 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?

An important issue for future work involves predicting an agent's behavior at revision states. Currently, we do not know how to model an agent's belief revision process and cannot predict an agent's action after an unexpected observation. Past actions do not tell how to ascribe belief in that case. We believe some form of an inductive leap is required, which should exploit additional structure, not present in our current model. Such structure could be obtained by e.g., augmenting our purely semantic construction with an interpretation of a suitable language over the possible states.[11]

This paper complements our previous work on belief ascription [Brafman and Tennenholtz, 1994b] and supplies initial answers to the above-mentioned questions. In this paper, we reviewed our proposed structure for 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 modeling.[12]

Acknowledgment: We thank Yoav Shoham for useful discussions relating to this work. The first author was partially supported by ARPA and AFOSR through grants AF F 49620-94-1-0090 and AF F49620-92-J-0547.

References

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

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

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

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

[Brafman, 1995] R. I. Brafman. *Mental State as a Modeling Tool: Theory and Applications.* PhD thesis, Stanford University, 1995. To appear 10/95.

[del Val and Shoham, 1993] Alvaro del Val and Yoav Shoham. Deriving properties of belief update from theories of action (II). In *IJCAP93, Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence,* pages 732-737, 1993.

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

[Friedman and Halpern, 1994] N. Friedman and J. Y. Halpern. 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, 1994. Morgan Kaufmann.

[Goldszmidt and Pearl, 1992] M. Goldszmidt and J. Pearl. 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),* pages 661-672, 1992.

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

[Katsuno and Mendelzon, 199l] H. Katsuno and A. Mendelzon. On the difference between updating a knowledge base and revising it. In *Principles of Knowledge Representation and Reasoning: Proc. Second Intl. Conf (KR '91),* pages 387-394, 1991.

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

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

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

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

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

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

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

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

[11] This point has been suggested to us by Hector Levesque.

[12] Additional results, comparison to work on plan recognition, and discussion of our general approach can be found in [Brafman, 1995].