ELSEVIER

Artificial Intelligence 94 (1997) 217-268

Artificial Intelligence

Modeling agents as qualitative decision makers

Ronen I. Brafman [a,*], Moshe Tennenholtz [b]

[a] *Department of Computer Science, University of British Columbia, Vancouver, B.C., Canada V6T 1Z4*

[b] *Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Haifa 32000, Israel*

Abstract

We investigate the semantic foundations of a method for modeling agents as entities with a mental state which was suggested by McCarthy and by Newell. Our goals are to formalize this modeling approach and its semantics, to understand the theoretical and practical issues that it raises, and to address some of them. In particular, this requires specifying the model's parameters and how these parameters are to be assigned (i.e., their *grounding*). We propose a basic model in which the agent is viewed as a qualitative decision maker with beliefs, preferences, and a decision strategy; and we show how these components would determine the agent's behavior. We ground this model in the agent's interaction with the world, namely, in its actions. This is done by viewing model construction as a constraint satisfaction problem in which we search for a model consistent with the agent's behavior and with our general background knowledge. In addition, we investigate the conditions under which a mental state model exists, characterizing a class of "goal-seeking" agents that can be modeled in this manner; and we suggest two criteria for choosing between consistent models, showing conditions under which they lead to a unique choice of model. © 1997 Elsevier Science B.V.

Keywords: Agent modeling; Mental states; Qualitative decision making; Belief ascription; Multi-agent systems; Prediction

1. Introduction

This article investigates the semantic foundations of a modeling method that uses formal notions of mental state to represent and reason about agents. In this method, agents are described *as if* they are qualitative decision makers with a mental state

* Corresponding author. Email: brafman@cs.ubc.ca.

[1] Email: moshet@ie.technion.ac.il.

PII S0004-3702(97)00024-6

consisting of mental attributes such as beliefs, knowledge, and preferences. The use of such models, which we refer to as *mental-level models*, was proposed by McCarthy [43] and by Newell [45], and our goals are to provide a formal semantic account of this modeling process, to understand some of the key issues it raises, and to address some of them.

The ability to model agents is useful in many settings. In particular, in multi-agent systems, the success of one's action or plan depends on the actions of other agents, and a good model of these agents can help construct more informed plans that are more likely to succeed. Mental-level models bring two promising properties to this task: they provide an abstract, implementation-independent way of representing agents, and they are built from intuitive and useful attributes, such as beliefs, goals and intentions. The abstract nature of mental-level models has a number of important practical implications.

(1) A single formalism can capture different agents running on different hardware platforms and written by designers who have used different programming languages and who have followed different design paradigms.
(2) We may be able to construct mental-level models without privileged access to the internal state of the agent because implementation details do not appear in these abstract models.
(3) Fewer lower-level details may mean faster computation.

The abstract nature of mental-level models is also important for theoretical analysis. It provides a uniform basis for comparing and analyzing agents, much like Levesque's *Computers as believers* paradigm [40] allows for abstract analysis of knowledge representation schemes. The second property, intuitiveness, is valuable in design validation since one approach to design validation is to transform low-level descriptions of agents that are difficult to analyze, such as procedural programs or mechanical designs, into intuitive high-level models. In addition, intuitive descriptions of agents in terms of their beliefs and goals are useful when we want to help these agents achieve their goals or correct erroneous beliefs. These abilities are sought after in cooperative multi-agent systems, in intelligent information systems, and in user interfaces, to name but a few areas.

1.1. Issues in mental-level modeling

Despite their promising properties, mental-level modeling has not been studied extensively in AI. The scarcity of citations on this issue in a recent survey of work on mental states within AI by Shoham and Cousins [58] attests to this fact.[2] Similarly, although Newell's paper on the *Knowledge Level* [45] is among the most referenced AI papers [5], in his perspective paper [46], Newell laments the lack of work following up on these ideas by the "logicist community". He mentions Levesque's [40] as the only exception.

Given this situation, it is worth clarifying what we view as the four central questions in mental-level modeling. They are:

[2] The only modeling related works there deal with plan recognition.

(1) Structure—what class of models should we consider?
(2) Grounding—how can we base the model construction process on a definite and objective manifestation of the agent?
(3) Existence—under what conditions will a model exist?
(4) Choice—how do we choose between different models that are consistent with our data?

The importance of the first question is obvious, however, the others deserve a few words of explanation.

Many researchers support an agent design approach in which the designer specifies an initial database of beliefs, goals, intentions, etc., which is then explicitly manipulated by the agent (e.g., [6,49,52,57] and much of the work on belief revision). Grounding may not seem a crucial issue for such work because human designers are likely to find mental attitudes natural to specify. However, grounding *is* crucial for modeling applications. The whole point here is that we cannot directly observe the mental state of another agent. Moreover, there are good reasons why general background knowledge alone will not do. First, there is no reason we should know the mental state of agents designed by other designers, a common case in multi-agent systems. Second, one cannot always predict into the distant future the mental state of agents that learn and adapt, even if she designed these agents. Finally, and perhaps most importantly, what we mean by the mental state of an agent is not even clear when this agent is not designed using a knowledge-based approach, for example, when it is a *C* program or the result of training a neural net. This last point is crucial if we take seriously Newell's idea of mental state models as abstract descriptions of agents. Grounding is important semantically even from the design perspective: it makes concrete the abstract Kripke semantics [35] that is often used in the literature, and it allows us to answer a central question in the theoretical analysis of agents and their design: Does program *X* implement mental-level specification *Y*?

While grounding has been discussed by some authors (see Section 8), the questions of model choice, and in particular, model existence have not received much attention. We see model existence as the central theoretical question in this area. Answers to it will allow us to evaluate any proposal for mental-level models by telling us under what conditions it is applicable and hence, what assumptions or biases we are making when we model agents in this manner. Techniques for choosing among possible models are crucial for practical applications, especially prediction, since different models may give rise to different predictions.

1.2. An overview of our approach

Having described the main questions in mental-level modeling, we proceed with an overview of this paper and its view of mental-level modeling.

Model structure

We propose a structure for mental-level models (Sections 2 and 3) that is motivated by work in decision theory [41] and previous work on knowledge ascription [20,54] in which the agent is described as a qualitative decision maker. This model contains

three key components: beliefs, preferences, and a decision criterion. We see these as the essential components of any mental-level structure, accounting for the agent's perception of the world, its goals, and its method of choosing actions under uncertainty.

The beliefs of the agent determine which among the possible states of the world it considers to be plausible. For example, the possible worlds of interest may be *rainy* and *not-rainy*, and the agent believes *rainy* to be plausible. The agent's preferences tell us how much it likes each outcome. For example, suppose the agent has two possible actions, taking or leaving an umbrella, whose outcomes are described by the following table:

	rainy	*not-rainy*
take umbrella	dry, heavy	dry, heavy, look stupid
leave umbrella	wet, light	dry, light

The agent's preferences tell us how much it values each of these outcomes. We will use real numbers to describe these values, where larger numbers indicate better outcomes:

	rainy	*not-rainy*
take umbrella	5	−1
leave umbrella	−4	10

The agent chooses its action by applying its decision criterion to the outcome of the different actions at the plausible worlds. A simple example of a decision criterion is *maximin*, in which the action with the best worst-case outcome is chosen. For example, if the agent believes both worlds to be plausible and uses *maximin*, it will choose the action *take umbrella*, since its worst case outcome is -1, which is better than the worst case outcome of *leave umbrella* (-4). However, if the agent believes only *not-rainy* to be plausible, it will choose *leave umbrella*, whose (plausible) worst case outcome (10) is now much better than that of *take umbrella* (-1).

Our description evolves in two stages: First, we model simple agents that take one-shot actions (Section 2); then, we extend these ideas to cover dynamic agents that can take a number of consecutive actions and make observations in between (Section 3). In the dynamic case, we must also stipulate a relationship between an agent's mental state at different times; we are especially concerned with belief change.

Model grounding

In Section 2.3, we show how a model for an agent can be constructed using the main semantic observation of this paper: mental attitudes receive their meaning in a context of other mental attitudes, and this whole context should be grounded in the behavior of the agent which is (in principle) observable. The main tool used here is the *agency hypothesis*, the hypothesis that

(1) the agent can be described via beliefs, preferences, and a decision criterion, as in the mental-level model outlined above and

(2) its ascribed mental state is related to its observed behavior in the specified manner.

Under this hypothesis, we can view the problem of ascribing a mental state to the agent as a constraint satisfaction problem. The agent's ascribed mental state must be such that

(1) it would have generated the observed behavior, and

(2) it is consistent with the background knowledge.

For instance, consider the above example, and suppose we have as background knowledge the agent's preferences (as specified in the table) and its decision criterion, which is *maximin*. If we observe the agent go out without an umbrella, we must conclude that it believes *not-rainy*, for if it had other beliefs, it would have taken a different action. We put special focus on this class of *belief ascription* problems.

Once a model of an agent has been constructed, it can be used to predict its future behavior. In Section 4, we pause to provide some insight on how the above ideas can be used to predict the behavior of the modeled agents based on past behavior. We exploit our ability to construct mental-level models of agents based on their behavior together with the stipulated relationship between the modeled agent's mental state and its next action.

Model existence and model choice

In Section 5, we examine model selection and suggest a number of criteria for choosing among models. One criterion prefers models that make fewer assumptions about the agent's beliefs, while the other criterion prefers models that (in some precise sense) have a better explanatory power. Then, we provide conditions under which these criteria lead to a unique model choice when a model exists. Next, in Section 6, we characterize a class of "goal-seeking" agents for whom a mental-level model can be constructed. These are agents whose behavior satisfies two rationality postulates.

Our work makes a number of contributions to the study of mental states in AI. First, we believe this is the first attempt to formalize this modeling approach. Hence, this paper enhances our understanding of this area and the main issues it involves. Second, we make a number of semantic contributions in addressing these issues.

(1) We make the notion of a decision criterion, which handles the issue of choice under uncertainty, an explicit part of the model.

(2) We provide grounding for mental-level models and mental attitudes.

(3) We are the first to emphasize and treat the question of model existence.

(4) We suggest two criteria for model choice and give conditions under which they lead to unique model choice.

As can be seen, our focus in this paper is on semantic issues. The concepts, structures, and processes we discuss should serve to deepen our understanding of these issues. They are not meant to be practical or implemented directly in their extensive form, although we hope that the insight gained can form the basis for appropriate algorithms and data structures.[3]

[3] As an analogy, naive use of models of first-order logic for the purpose of logical deduction is not a very good idea. Instead, one would either use a theorem prover (which manipulates symbols, not models) or some model checking method employing an efficient encoding of models.

Our work builds on much previous work; In Section 7 we discuss its relationship with work in the areas of economics and game theory, and in Section 8 we analyze the relationship of this work with existing work in AI. We conclude with a short discussion of implementation issues and future work in Section 9.

2. Static mental-level models

This section provides a formal account of a proposed structure for mental-level models and their ascription process. Various parts of the model we propose should be familiar to readers acquainted with decision theory, game theory, and in particular, qualitative decision making techniques (see, e.g., [41]). We examine static agents that take only one-shot actions (dynamic agents are discussed in Section 3). These agents enter one of a possible set of states, perform an action and restart. We start in Section 2.1 with a motivating example introducing the central concepts used in our model. The model itself, in which the mental state of the agent is described in terms of its beliefs, preferences, and a decision criterion, is described in Section 2.2. This mental state is then related and grounded in the agent's behavior in Section 2.3 which examines the problem of ascribing a mental state to an agent, and, in particular, ascribing beliefs.

2.1. Motivating example

In order to help the reader relate to the more formal treatment that follows, we shall start with an example that introduces the central concepts encountered later.

Example 1. Consider the problem of designing an automated driver. This complex task is made even more complex if our driver is to exercise defensive driving. Defensive driving requires the driver to reason about the behavior of other drivers in order to predict their future behavior. These predictions, in turn, are used by the driver in order to choose actions that promote its safety.[4] We shall use a very simple version of this problem to illustrate some of our ideas.

The modeled agent, A, is another vehicle observed by our agent approaching an intersection with no stop sign. Our agent is approaching this intersection as well, and it has the right of way. It must decide whether or not it should stop, and in order to do so, it will try to determine whether or not agent A will stop. Its first step is to construct a model of agent A's state. It constructs this model by using its information about agent A's behavior together with general background information about drivers. Our agent models agent A as a qualitative decision maker using the model whose fundamental components are shown in Fig. 1.

[4] This scenario is motivated by work on fighter pilot simulators which are used in air-combat training. Such simulators must perform similar reasoning in the context of even more complex group activity. The ability to reason about the mental state of opponents, i.e., their beliefs, goals, and intentions, is important for building good simulators. Groups in Australia (AAII) and California (ISI) have incorporated such technology in their commercial systems to a limited extent [61].

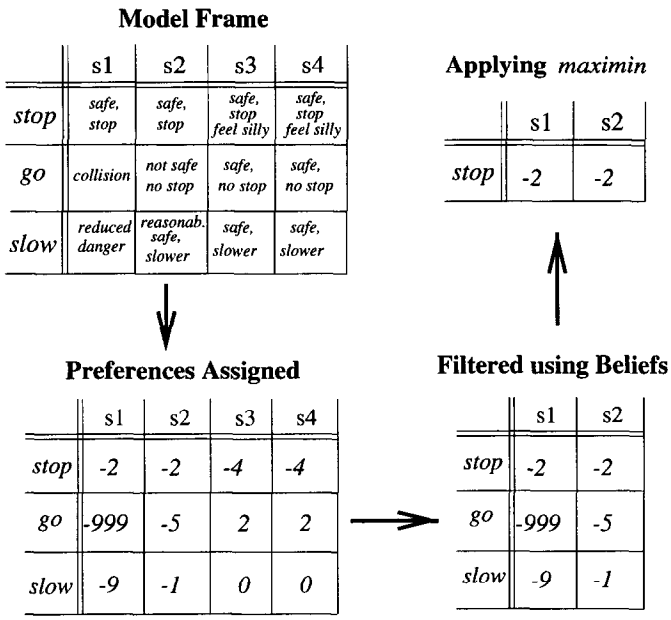


Fig. 1. Structure of mental-level models.

First, the modeler supplies a context, or *model frame*, by choosing a relevant set of possible states of the world and a set of actions and their outcomes. Our agent shall consider the following states as possible for agent A:

$s1$—another car is crossing and it has right of way;

$s2$—another car is crossing and A has right of way;

$s3$—no other car is crossing and A has right of way;

$s4$—no other car is crossing and A does not have right of way.

The set of possible actions is: {*stop*, *continue*, *slow-down*} (abbreviated as *stop, go, slow* in Fig. 1). Each action leads to a particular outcome when executed in one of the possible worlds. For instance, if *continue* is executed at state $s1$, a collision will follow. Fig. 1 (top-left table) contains a more complete description of these outcomes.

Next, our agent must assess the desirability of each outcome from agent A's point of view. Naturally, collision is a pretty bad outcome; stopping when there is no car is also not a good option, though much better than collision. It is often convenient to specify preferences numerically, by means of a *value* function, such that lower values are assigned to less desirable outcomes. In our context, we would expect our agent to be able to reasonably assess agent A's value function. Fig. 1 (bottom-left table) provides one reasonable choice for this function.

Presumably, agent A has some information about the actual state of the world. Let us suppose for the moment that its information indicates that $s1$ and $s2$ are plausible. When agent A decides which action to take, it will consider the effects of each action on each of the plausible worlds—we call this the *plausible outcome* of the action—and

it will compare these effects in order to choose the action that gives rise to the most preferred plausible outcome.

Examining the bottom-right table in Fig. 1, we see that action *stop* is better when $s1$ is the case, while *slow* is better when $s2$ is the case; hence, the choice between them is not obvious. This is where agent A's *decision criterion* comes in. A decision criterion is simply a method for comparing the values of alternative actions in the context of a number of plausible states. Let us suppose that our agent assumes that agent A is cautious and employs the *maximin* criterion mentioned earlier. In that case, it will predict that agent A will stop.

Unfortunately, our agent does not have access to the internal state of agent A, so it cannot observe A's "beliefs". Therefore, in order to complete the ascription process, it must be able to ascribe beliefs to A using its information about A and its general information about drivers. We refer to this as the *belief ascription* problem. The main criteria that the ascribed model must satisfy is descriptive accuracy. Hence, the ascribed beliefs of agent A (in fact, the whole ascribed model) must be such that it would predict the actions that were actually observed. Thus, belief ascription is a special instance of a constraint satisfaction problem, where the constraints are given by the stipulated relationship between the various components of the model and our agent's observations.

In our particular scenario, our agent observes that agent A continues driving without slowing down. The sets of plausible worlds that agent A could have that would lead it to take this action are: $\{s3\}, \{s4\}, \{s3, s4\}$. Notice that our agent does not have a unique belief ascription for agent A, i.e., there are a number of models of agent A that are consistent with our agent's information. In Section 5 we shall consider various domain-independent heuristics that can be used to choose among a set of models consistent with our information.

We have just seen an example of belief ascription. Our agent has concluded that agent A does not "believe" there is another vehicle approaching the intersection since agent A's (ascribed) set of plausible worlds cannot contain states s_1 or s_2. This conclusion is useful if we wish to improve agent A's design, as it points out a flaw in its reasoning, sensing, or decision making abilities. Otherwise, a model of an agent's beliefs may not be very useful in the one-shot static model we describe in this section. However, it will become useful in the dynamic setting discussed later, where our agent could use it to predict agent A's future behavior. Intuitively, subject to the assumption that A's beliefs persist until new contradictory information arrives, our agent can predict that agent A will continue at its current speed. Consequently, our agent must stop in order to prevent an accident.

2.2. Model structure

In this section, we describe a formal model of static agents. At this stage, we shall not concern ourselves with how one constructs a model for a particular agent but rather with the question of what type of structure should be used to describe mental states. Many of the definitions used here are standard and should be familiar to readers acquainted with formal work on mental states and the fundamentals of decision theory.

We start with a low-level description of an agent, motivated by the work of Halpern and Moses [26] and of Rosenschein [54], on top of which the mental-level model is defined. To clarify our definitions, we will accompany them with a simplified version of McCarthy's thermostat example [43]. The choice of modeling a thermostat, which we normally do not view as having beliefs, stresses our view of mental states as modeling abstractions.

Example 2. In [43], McCarthy shows how we often ascribe mental states to simple devices. Our goal is to formalize his informal discussion of thermostats. We assume that we have a thermostat in a room that controls the flow of hot water into that room's radiator. The thermostat can either *turn-on* or *shut-off* the hot water supply to this radiator. It chooses its action based on whether it senses the temperature of the room to be above or below a certain threshold value.

Describing agents. An *agent* is described as a state machine, having a set of possible (local) states, a set of possible actions, and a program, which we call its *protocol*. The agent functions within an *environment*, also modeled as a state machine with a corresponding set of possible states. Intuitively, the environment describes all things external to the agent, including possibly other agents. We refer to the state of the whole 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, and that the effects of the agent's actions are a (deterministic) function of its state and the environment's state.[5] These effects are described by the *transition function*. Thus, the agent and the environment can be viewed as 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, and those global states that are possible are called *possible worlds*.

Definition 3. An *agent* is a three-tuple $\mathcal{A} = \langle L_{\mathcal{A}}, A_{\mathcal{A}}, \mathcal{P}_{\mathcal{A}} \rangle$, where $L_{\mathcal{A}}$ is the agent's set of *local states*, $A_{\mathcal{A}}$ is its set of *actions*, and $\mathcal{P}_{\mathcal{A}} : L_{\mathcal{A}} \rightarrow A_{\mathcal{A}}$ is its *protocol*. $L_{\mathcal{E}}$ 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* is a subset S of the set of global states $L_{\mathcal{A}} \times L_{\mathcal{E}}$. Finally, the *transition function*, $\tau : (L_{\mathcal{A}} \times L_{\mathcal{E}}) \times A_{\mathcal{A}} \rightarrow (L_{\mathcal{A}} \times L_{\mathcal{E}})$ maps a global state and an action to a new global state.

In the sequel we often consider partial protocols in which an action is not assigned to each state. Partial protocols allow us to represent information about an agent whose behavior has been observed only in a subset of its possible states.

Example 2 (*continued*). For our thermostat, $L_{\mathcal{A}} = \{-, +\}$. The symbol $-$ corresponds to the case when the thermostat indicates a temperature that is less than the desired

[5] A framework in which the environment does act and in which the outcomes of actions are non-deterministic can be mapped into our framework using richer state descriptions and larger sets of states, a common practice in game theory.

room temperature, and the symbol + corresponds to a temperature greater or equal to the desired room temperature. The thermostat's actions, $A_{\mathcal{A}}$, are $\{turn\text{-}on, shut\text{-}off\}$. The environment's states, $L_{\mathcal{E}}$, are $\{cold, ok, hot\}$. We do not assume any necessary relation between the states of the thermostat and the environment, taking into account the possibility of measurement error. Therefore, the set of possible worlds is exactly $L_{\mathcal{A}} \times L_{\mathcal{E}}$. We chose the following transition function:

	$(-, cold)$	$(+, cold)$	$(-, ok)$	$(+, ok)$	$(-, hot)$	$(+, hot)$
turn-on	$(-, ok)$	$(+, ok)$	$(-, hot)$	$(+, hot)$	$(-, hot)$	$(+, hot)$
shut-off	$(-, cold)$	$(+, cold)$	$(-, ok)$	$(+, ok)$	$(-, ok)$	$(+, ok)$

In this example, the effects of an action on the environment do not depend on the state of the thermostat. In addition, the fact that we use a static, "one-shot" model allows us to make certain simplifications. First, we do not explicitly model external influences on the room's temperature other than those stemming from the thermostat's actions. Second, we ignore the effect of the thermostat's actions on its state since it is inconsequential—we adopt the convention that this state does not change.

Finally, the thermostat's protocol is the following:

state	−	+
action	turn-on	shut-off

Given the set S of possible worlds, we can associate with each local state l of the agent a subset of S, $PW(l)$, consisting of all worlds in which the local state of the agent is l.

Definition 4. The agent's set of *worlds possible at* l, $PW(l)$, is defined as $\{w \in S \mid$ the agent's local state in w is $l\}$.

Thus, $PW(l)$ are the worlds consistent with the agent's state of information when its local state is l. Halpern and Moses [26] and Rosenschein [54] use this definition to ascribe knowledge to an agent at a world w. Roughly, they say that the agent *knows* some fact φ at a world w if its local state l in w is such that φ holds in all the worlds in $PW(l)$.

Example 2 (*continued*). While the thermostat, by definition, knows its local state, it knows nothing about the room temperature. Formally, this lack of knowledge follows from the fact that we made all elements of $L_{\mathcal{A}} \times L_{\mathcal{E}}$ possible, e.g., $(-, hot)$ is a possible world. Intuitively, we are allowing for the possibility of a measurement error by the thermostat.

Example 1 (*continued*). Let us try to relate these first definitions to the modeling problem faced by our driving agent. Agent A, which our agent tried to model, may be quite complex. Yet, our agent cared about modeling A in its current circumstances only

(i.e., as it is approaching the intersection). Consequently, our agent cared about A's current local state (call it l), A's current possible actions (i.e.,$\{stop, continue, slow\text{-}down\}$), and A's current action (i.e., *continue*). The possible states of the environment are s_1, s_2, s_3, s_4, and so the possible global states are $(l, s_1), (l, s_2), (l, s_3), (l, s_4)$. The transition function is described implicitly in Fig. 1.[6]

If truth assignments (for some given language) are attached to each world in S, and if a world s' is defined to be accessible from s whenever the agent's local states in s and s' are identical, we obtain the familiar $S5$ Kripke structure.

Belief. While knowledge (or $PW(l)$) defines what is theoretically possible, belief defines what, in the eyes of the agent, is the set of worlds that should be taken into consideration. We describe the agent's beliefs using a *belief assignment*, a function that assigns to each local state l a nonempty subset of the set of possible worlds. Beliefs are modeled as a function of the agent's local state because this state provides a complete description of the agent at a point in time, so it should also determine its beliefs. The role beliefs play is to divide the worlds possible at a local state to those that are *plausible*, and thus, are worthy of consideration, and those that are not plausible, and can be ignored.

Definition 5. A *belief assignment* is a function, $B : L_{\mathcal{A}} \rightarrow 2^S \setminus \emptyset$, such that for all $l \in \mathcal{L}_{\mathcal{A}}$ we have that $B(l) \subseteq PW(l)$. We refer to $B(l)$ as the worlds *plausible* at l.

Example 2 (*continued*). One possible belief assignment, which would probably make the thermostat's designer happy, is $B(-) = \{-, cold\}$ and $B(+) = \{+, hot\}$. From now on we will ignore the agent's local state in the description of the global state and write, e.g., $B(+) = \{hot\}$.

We remark that (after adding interpretations to each world) this approach yields a $KD45$ belief operator and a relationship between knowledge and belief that was proposed by Kraus and Lehmann in [31].

Preferences. Beliefs really make sense as part of a fuller description of the agent's mental state, which has additional aspects. One of these aspects is the agent's preference order over possible worlds, which may be viewed as an embodiment of the agent's desires. There are various assumptions that can be made about the structure of the agent's preferences. In most of this section, we will only assume that they define a total pre-order on the set of possible worlds S. However, in some cases, we may need a richer algebraic structure, e.g., one in which addition is defined (e.g., for the principle of indifference). In what follows we will use *value functions* to represent the agent's preferences.

[6] To precisely conform with the definitions, we would have had to include the set of outcomes in the set of possible states and to define the effects of actions on these outcomes.

Definition 6. A *value function* is a function $u : S \rightarrow \mathbb{R}$.

This numeric representation is most convenient for representing preferences. Under this representation, the state s_1 is at least as preferred as state s_2 *iff* $u(s_1) \geqslant u(s_2)$.

Value functions are usually associated with the work of von Neumann and Morgenstern on utility functions [62]. However, their utility functions express more than a simple pre-order over the set of states, and we do not need to incorporate all of their additional assumptions.

Because this section is concerned with simple agents that take one-shot actions and restart, we can view values as a function of state. In Section 3, we will need to look at more complex value functions that take into account sequences of states rather than single states.

Example 2 (*continued*). The goal of our thermostat is for the room temperature to be *ok*. Thus, it prefers any global state in which the environment's state is *ok* over any global state in which the environment's state is either *cold* or *hot*, and is indifferent between *cold* and *hot*. In addition, it is indifferent between states in which the environment's state is identical, e.g., $(-, ok)$ and $(+, ok)$. This preference order can be represented by a value function which assigns 0 to global states in which the environment's state (i.e., the room temperature) is *hot* or *cold*, and which assigns 1 to those states in which the environment's state is *ok*.

Plausible outcomes. When the exact state of the world is known, the result of following some protocol, $\mathcal{P}$, is also precisely known. (Remember that actions have deterministic effects). Therefore, we can evaluate a protocol by looking at the value of the state it would generate in the actual world. However, the more common situation is that the agent is uncertain about the state of the world and considers a number of states to be plausible. Then, we can represent its view of how desirable a protocol $\mathcal{P}$ is in a local state l by a vector whose elements are the values of the plausible states $\mathcal{P}$ generates, i.e., the worlds generated by using $\mathcal{P}$ at $B(l)$. We refer to this tuple as the *plausible outcome* of $\mathcal{P}$.

Example 2 (*continued*). The following table gives the value of the outcome of each of the thermostat's possible actions at each of the environment's possible states (where $\star$ stands for either $-$ or $+$):

	$(\star, cold)$	$(\star, ok)$	$(\star, hot)$
turn-on	1	0	0
shut-off	0	1	1

If the thermostat "knew" the precise state of the world, it would have no trouble choosing an action based on the value of its outcome. For example, if the state is *cold*, *turn-on* would lead to the best outcome. When there is uncertainty, the thermostat must compare *vectors* of plausible outcomes instead of single outcomes. For example,

if $B(l) = \{cold, ok\}$, the plausible outcome of the action *turn-on* is $(1,0)$, and the plausible outcome of the action *shut-off* is $(0,1)$.

Definition 7. Given a transition function τ, a belief assignment B, and an arbitrary, fixed enumeration of the elements of $B(l)$, the *plausible outcome* of a protocol $\mathcal{P}$ in l is a tuple whose kth element is the value of the state generated by applying $\mathcal{P}$ starting at the kth state of $B(l)$.

Note that because we are considering only static agents, we could have spoken about plausible outcomes of actions instead of plausible outcomes of protocols. For static agents both are identical.

Decision criteria. While values are easily compared, it is not a priori clear how to compare plausible outcomes, and thus, how to choose among protocols. A strategy for choice under uncertainty is required, which depends on, e.g., the agent's attitude towards risk. This strategy is represented by the *decision criterion*, a function taking a set of plausible outcomes, returning the *set* of most preferred among them.

We have previously encountered the *maximin* criterion, which selects those tuples whose worst case outcome is maximal.[7] Another example is the *principle of indifference* which selects those tuples whose average outcome is maximal.[8] (For a fuller discussion of decision criteria consult [41].)

Example 8. Consider the example given in the introduction in which the following matrix was used:

	rainy	*not-rainy*
take umbrella	5	−1
leave umbrella	−4	10

We have seen that when both worlds are plausible the two plausible outcomes are $(5,-1)$ and $(-4,10)$. When the *maximin* criterion is used, the first one, corresponding to *take umbrella* is the most preferred. However, when the *principle of indifference* is used, the plausible outcome of *leave umbrella* is preferred.

Definition 9. A *decision criterion* is a function $\rho : \bigcup_{n\in\mathbb{N}} 2^{\mathbb{R}^n} \to \bigcup_{n\in\mathbb{N}} 2^{\mathbb{R}^n} \setminus \emptyset$ (i.e. from/to sets of equal length tuples of reals), such that for all $\mathcal{U} \in \bigcup_{n\in\mathbb{N}} 2^{\mathbb{R}^n}$ we have that $\rho(\mathcal{U}) \subseteq \mathcal{U}$ (i.e., it returns a non-empty subset of the argument set).

Notice that we can use a decision criterion to compare tuples. For instance, if $\rho(\{w,v\}) = \{w\}$ then we can say that w is more preferred than v.

[7] We apply this criterion recursively, i.e., when two tuples have the same worst case outcome, we compare the next to worst, and so on.

[8] With an infinite set of tuples, *maximin* and the *principle of indifference* may not have a set of most preferred tuples and appropriate variants of these criteria must be defined.

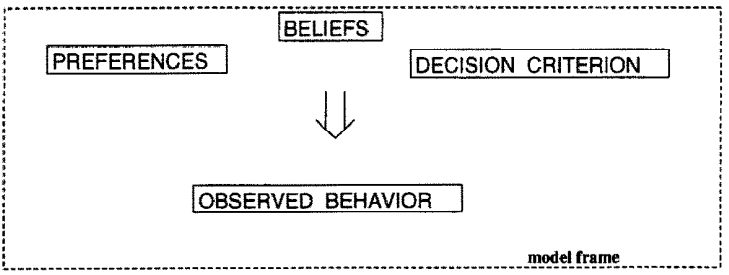


Fig. 2. The agency hypothesis.

The agency hypothesis. We can capture the relationship among the various components of our mental-level model using the following:

Definition 10. The *agency hypothesis*: The agent's actual protocol has a plausible outcome that is most preferred (according to the agent's decision criterion) among the set of plausible outcomes of all possible protocols.[9]

The agency hypothesis takes the view of a rational balance among the agent's beliefs, values, decision criterion and behavior (see Fig. 2). It states that the agent chooses actions whose plausible outcome is maximal according to its decision criterion. Thus, the choice of the protocol is dependent upon $B(l)$ and u, which define the plausible outcome of each protocol, and ρ, which chooses among these different plausible outcomes. By viewing the agent as a qualitative decision maker, the agency hypothesis attributes some minimal rationality to the agent, assuming that it would take an action whose plausible outcome is most preferred according to some decision criterion.

We can now formally define a notion of a *mental-level* model.

Definition 11. A *mental-level* model for an agent $\mathcal{A} = \langle \mathcal{L}_{\mathcal{A}}, A_{\mathcal{A}}, \mathcal{P}_{\mathcal{A}} \rangle$ is a tuple $\langle \mathcal{L}_{\mathcal{A}}, A_{\mathcal{A}}, B, u, \rho \rangle$, where B is a belief assignment, u is a value function, and ρ is a decision criterion.

Thus, a mental-level model provides an abstract implementation-independent description of the agent. Instead of describing its local state in terms of the values of various registers, or the values of state variables, we use implementation-independent notions: beliefs, preferences, and a decision criterion.

2.3. Model ascription

We have taken the first step towards formalizing mental-level models by proposing a general structure. Now, we come to the construction task, where we must explain the process by which one can model a *particular* agent, i.e., the process by which

[9] The agent's possible protocols, are implicitly defined by the set of actions $A_{\mathcal{A}}$ (cf. Definition 3).

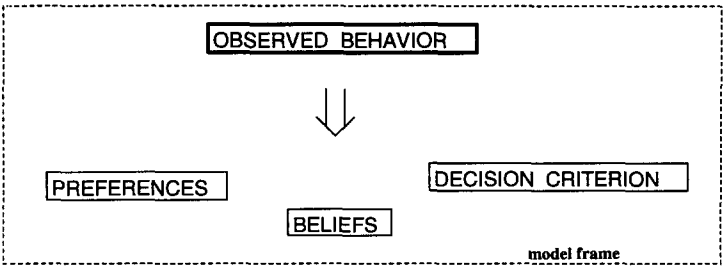


Fig. 3. Model construction.

we actually ascribe a mental state to some particular agent. This process should require information that we can realistically expect the modeling agent to have. This information should primarily consist of the modeled agent's behavior, which is a concrete, observable aspect of it. This behavior is formally captured by our notion of a protocol, or a partial protocol when we only know how the agent acts in certain situations. Thus, the modeled agent's ascribed mental state should be grounded in its *behavior*.

The agency hypothesis supplies a basis for this process by providing constraints on the agent's mental state given its behavior. That is, it makes only certain mental states consistent with a given behavior: those mental states that would have induced such behavior.

Definition 12. A mental-level model for $\mathcal{A}$, $\langle \mathcal{L}_{\mathcal{A}}, A_{\mathcal{A}}, B, u, \rho \rangle$ is *consistent* with a protocol $\mathcal{P}$ if for all $l \in \mathcal{L}_{\mathcal{A}}$ it is the case that the plausible outcome of $\mathcal{P}$ is most preferred according to B, u, and ρ.

It is consistent with a partial protocol $\mathcal{P}'$ if it is consistent with some completion $\mathcal{P}$ of $\mathcal{P}'$.

This key definition tells us that a mental-level model is consistent with our observations of the agent when the model is such that an agent with this mental state would display the observed behavior. This definition embodies two key ideas about the semantics of mental states:

(1) The agent's *ascribed* mental state is grounded in its behavior (formally captured by its protocol).
(2) Separate mental attitudes, such as belief and preference, are not interpreted by themselves, but rather, they receive their meaning in the context of the agent's whole mental state (see Fig. 3).

In some applications, such as design validation and program analysis, we can "play around" with the modeled entity and obtain sufficient observations to constrain the class of consistent models. However, in other applications, such as agent modeling within a multi-agent system, we may not be able to make enough observations to reduce the class of consistent models to a manageable size. Hence, we will have to introduce assumptions that reduce this class of models. The problem of ascribing a mental state to an agent under assumptions Ψ and based on a (possibly partial) protocol $\mathcal{P}$ can be stated as:

Find a model of the agent consistent with $\mathcal{P}$ that satisfies the assumptions Ψ.

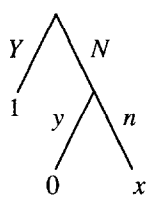
In this paper, we are particularly interested in a special class of constrained mental-level modeling problems, that of *belief ascription*. In belief ascription our task is to supply one component of a mental-level model, the beliefs, given assumptions about the other components: the preferences and the decision criterion. However, the general approach presented so far can be applied to other ascription problems, e.g., goal ascription. The belief ascription problem is formally defined as follows:

> Given an agent $\mathcal{A} = \langle \mathcal{L}_\mathcal{A}, A_\mathcal{A}, \mathcal{P}_\mathcal{A} \rangle$, a value function u, and the decision criterion ρ, find a belief assignment B such that $\langle \mathcal{L}_\mathcal{A}, A_\mathcal{A}, B, u, \rho \rangle$ is consistent with $\mathcal{P}_\mathcal{A}$.

Example 2 (*continued*). Given our knowledge of the thermostat, what beliefs can we ascribe to it? We know the thermostat's protocol and goals. We will assume that its decision criterion simply prefers tuples that are not dominated by any other tuple.[10] Given this, we have the following constraints on the thermostat's beliefs: $B(-) \supseteq \{cold\}$ and at least one of *ok* or *hot* are in $B(+)$. If the thermostat's beliefs violate these constraints, the plausible outcome of the action prescribed by its protocol would be strictly less preferred than the plausible outcome of the other action.

Our treatment so far can be viewed as assuming knowledge of the agent's local state. While access to such information is possible in certain contexts (e.g., when the modeling is done by the designer), it is not likely in many other contexts. Yet, notice that we did not make use of such information; the ascription process described here relied only on the modeled agent's action. In general, knowing the local state matters when the agent's possible worlds differ depending on the local state, e.g., when the agent has reliable sensors that allow it to rule out certain states of the environment as impossible. This task of determining the set of possible worlds is a difficult one, and we view it as part of the model framing problem.

Example 13 (*A simple game*). The following tree describes a one-person decision problem based on a game that appears in [34]:



Initially the agent decides whether to choose Y or N. If Y is chosen, a payoff of 1 is obtained, otherwise the environment chooses either y, with a payoff of 0 to the agent, or n, with a payoff of $x > 1$. While game theoreticians are mostly concerned with how games should be played when the environment is another rational agent, we ask a

[10] Tuple v dominates w if every component of the v is at least as good as the corresponding component of w and some component of v is strictly better than the corresponding component of w.

simple question: what can we say if we observed the agent's first move to be N? This question is interesting because it is easy to construct a two person game based on this decision problem in which N is not a "rational" move. Such behavior, while perhaps irrational in some sense, can still be understood as rational given certain beliefs, e.g., that the environment will play n.

The following payoff matrix describes the agent's decision problem (the different states of the world correspond to the environment's behavior if N is played):

	y	n
Y	1	1
N	0	x

Having chosen N, if the agent's decision criterion is *maximin* then regardless of the value of x, the agent must believe that the environment will play n. Belief that y is plausible is inconsistent with the agent's behavior, since it would imply that Y should be chosen.

In the case of the *principle of indifference*, if $x < 2$, N is chosen only if the agent believes n to be the only plausible world. If $x \geqslant 2$ then a belief that both worlds are plausible would also cause N to be preferred.

Another decision criterion is *minimax regret*. The regret of performing action *act* in a state s is the difference between the best that can be done in state s and the actual payoff of *act* in s. This decision criterion prefers actions whose maximal regret is minimal. Here is the "regret" matrix for our decision problem:

	y	n
Y	0	$x-1$
N	1	0

For an agent following *minimax regret*, if $x < 2$ the agent must believe n to follow N, otherwise it may believe either n or $\{n, y\}$.

The idea of ascribing to an agent those belief assignments that make it satisfy the agency hypothesis leads to an interesting semantics for belief, stemming from its grounding in actions: The agent's (ascribed) plausible worlds consist of *those states that affect its choice of action.*

To better understand this subtle point, consider the special case of an agent whose set of possible actions can be varied arbitrarily. We subject this agent to an experiment in which we examine its chosen action in a local state l given different sets of possible actions. A sufficient condition for a global state $s \in PW(l)$ to be considered plausible by this agent (i.e., $s \in B(l)$) is that a pair a, a' of actions exists, such that for all $s' \in PW(l) \setminus \{s\}$ we have that $\tau(s', a) = \tau(s', a')$ but $\tau(s, a) \neq \tau(s, a')$ and the agent would choose a over a'. That is, a and a' have identical effects on all the worlds in $PW(l)$ except s. Thus, the agent must consider s to be plausible. Otherwise, the plausible outcomes of a and a' would have been identical, and the agent

would have not shown preference for one action over the other. This view of beliefs is closely related to Savage's notion of null-states [56] and Morris' definition of belief [44].

3. Dynamic mental-level models

In the previous section, we described a model of simple static agents and the basis for its construction. In this section we consider a more complex, dynamic model of agents that can take sequences of actions interleaved with observations. While many of the basic ideas remain the same, some definitions must be adjusted, and a number of new issues arise, most significantly, the question of how mental states at different times relate to each other. Using a running example, we start by considering those aspects of the static model that are inadequate for modeling dynamic agents and show how they can be revised in light of our desire to capture the relationship between the agent's state at neighboring time points. Some aspects of this problem of state dynamics are considered in Section 3.1, where the dynamics of the agent's preferences are discussed. More central is the problem of modeling the dynamics of an agent's beliefs. This is the topic of Section 3.2 in which one approach is suggested. In Section 3.3, we examine the problem of ascribing a mental state to a dynamic agent, which is not an easy task. This task can be simplified if we are able to reduce it to multiple instances of the problem of mental state ascription in the static model, and we examine the conditions under which this is possible.

3.1. A view of dynamic agents

Consider a mobile robot acting within a grid world whose task is to visit three locations on its way from the initial position I to the goal position G. This domain and one possible path are depicted in Fig. 4. We shall assume that at each position the robot either stops or moves in one of four possible directions. The robot can start at each of the positions marked I (in the example it starts at I_0); it has a position sensor that is accurate to within a distance of one, i.e., the sensor may indicate the actual position or any adjacent position; and its local state consists of all its past position readings.

We could use a static model to capture this setting by viewing the agent as making a single choice among possible protocols (often called *policies* or *strategies*) in the initial state.[11] However, this representation is too coarse and does not adequately support the task of prediction. Intuitively, the process of prediction (more fully discussed in Section 4) requires modeling an agent at a particular time point and using this model to predict its behavior in the future. This, in turn, requires us to use a model that makes explicit the relationship between the state of the agent at different time points.

[11] This representation of the agent's decision problem is referred to as a *strategic form* description in the game theory literature.

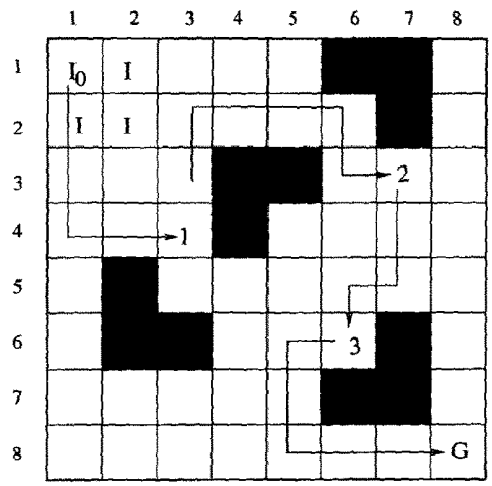


Fig. 4. A task requiring multiple actions.

In order to model the behavior described in Fig. 4 adequately, a number of obvious changes must be made to our model. First, we are no longer considering single states, but sequences of states, e.g., in the above example, we care about the robot's path. Following [20], we shall use the term *run* to refer to the sequence of global states of the agent/environment system starting with its initial state.

Runs describe the state of the agent and the environment over time. A run is *feasible* with respect to an agent $\mathcal{A}$ if it is a sequence of global states $s_0, s_1, \ldots$ such that for all $k > 0$, there exists some $a \in A_{\mathcal{A}}$ for which $\tau(s_{k-1}, a) = s_k$. A run is *protocol consistent* (*consistent* for short) with respect to an agent $\mathcal{A}$ if it is a feasible run $s_0, s_1, \ldots$ such that for all $k > 0$, $\tau(s_{k-1}, a) = s_k$ and a is consistent with $\mathcal{A}$'s (partial) protocol.[12] That is, this run could be generated by a particular sequence of actions of the agent that is consistent with its (possibly partial) protocol. We denote the set of all feasible runs by $\mathcal{R}$, and from now on, by a run we always mean a *feasible run*. We denote the set of suffixes of feasible runs by $\mathcal{R}_{suf}$. Intuitively, a modeling agent needs only consider consistent runs when it attempts to ascribe the modeled agent's beliefs about the past, having observed the modeled agent's past behavior. However, the modeling agent must consider all feasible runs as possible when it is attempting to predict the modeled agent's future behavior.

Next, we reconsider our definitions of beliefs and preferences. Now, our robot cares about runs rather than single states, and it is natural to describe it as having beliefs over run prefixes and preferences over run suffixes rather than over possible worlds. That is, at each state, it has beliefs concerning its current and past states, and it has preferences over its future states. Thus, when the robot is at position 2, it has various beliefs about what actual path it followed to its current state, and it has various preferences over

[12] That is, either a is assigned by $\mathcal{A}$'s partial protocol to its local state in s_{k-1}, or the protocol does not specify an action on this particular local state.

how it should proceed from here on. It may believe that its path consisted of the path depicted in Fig. 4 up to position 2, or it may consider a number of similar possible paths. Similarly, according to the task description, it would prefer run suffixes that will take it to position 3 and then to the goal over all other run suffixes. [13]

In order to formalize these ideas, we redefine the value function to be over the set of run suffixes.

Definition 14. A *value function* is a function $u : \mathcal{R}_{suf} \rightarrow \mathbb{R}$.

The definition of beliefs should change, too. Intuitively, instead of having a set of plausible worlds, the agent should have a set of plausible run prefixes. In order to simplify the model conceptually, we can represent run prefixes using their initial state (with the implicit assumption that the protocol, at least up to the current state, is known). This follows from our modeling assumption that actions have deterministic effects, and hence, each initial global state and protocol determine a unique run. [14] For example, in Fig. 4, the first 5 motion actions combined with the initial state (which cannot be completely depicted in the figure) determine a unique current global state. In this model, the environment's state in each global state describes the robot's actual position, as well as the readings it will receive if it were to reach a certain position at a certain time. The local state describes the sequence of position readings obtained so far.

Definition 15. Let $I \subseteq L_{\mathcal{A}} \times L_{\mathcal{E}}$ be the set of *initial worlds*, and define the function PW_I on local states as follows:

$$PW_I(l) = \{s_0 \in I \mid s_0, s_1, \ldots \text{ is a consistent run of } \mathcal{A}, \text{ and there exists some integer } n \text{ and environment state } e \text{ such that } s_n = (l, e)\}.$$

$PW_I(l)$ is called the set of *possible initial worlds* at l.

A *belief assignment* is a function $B_I : L \rightarrow 2^I \setminus \emptyset$ such that for all $l \in L$ we have that $B_I(l) \subseteq PW_I(l)$.

$PW_I(l)$ tells us from which initial global states it is possible to reach the local state l.

Let us consider how the robot's beliefs evolve. Suppose that among its possible initial worlds the robot finds those in which the position readings are accurate to be the most plausible. Hence, initially it will have four plausible initial worlds, corresponding to the four possible initial positions; each of these worlds will embody the assumption that future sensor readings will be correct. Suppose that the robot starts sensing only after its first motion command, which is "go down", and it receives a reading of $(2,2)$ as its current position. Consequently, it will believe that its current position is $(2,2)$.

[13] In certain situations, preferences over whole runs may be more appropriate. For example, depending on a number of modeling choices, this may be the case when the agent has no information about its past positions. We refer the reader to [8] for a more complete discussion of this issue.

[14] As we remarked in Section 2, non-determinism is handled by transforming all uncertainty about the effect of actions to uncertainty about the initial state of the environment. This means that the initial state specifies what position reading the robot will obtain at each point in time at each possible position.

Clearly, this belief is equivalent to the belief that its initial position was $(1,2)$. Indeed, sometimes we find it more convenient to represent the agent's beliefs using a set of plausible current worlds, rather than a set of plausible initial worlds.

Definition 16. $PW_{cur} : L \rightarrow 2^S$ assigns to each local state l its set of *possible current worlds*:

$$PW_{cur}(l) = \{s_n \mid s_0, s_1, \ldots \text{ is a consistent run of } \mathcal{A},\ s_0 \in I, \text{ and there exists some environment state } e \text{ such that } s_n = (l, e)\}.$$

The *current belief assignment* is the function $B_{cur} : L \rightarrow 2^S \setminus \emptyset$ such that

$$B_{cur}(l) = \{s_n \mid s_0, s_1, \ldots \text{ is a consistent run of } \mathcal{A},\ s_0 \in B_I(l), \text{ and for some integer } n \text{ and environment state } e \text{ we have that } s_n = (l, e)\}.$$

That is, $PW_{cur}(l)$ will contain a possible world s if the agent's local state in s is l, and s appears in a run which commences at one of the initial worlds. $B_{cur}(l)$ will contain those currently possible worlds occurring in a consistent run that commences in a plausible initial world.

The plausible outcome of a protocol $\mathcal{P}$ at a local state l is defined much as before. For our robot, given a set of plausible current states, each protocol would define a set of run suffixes which correspond to some path with associated sensor readings. Each such path has some value, as we explained earlier. Hence, with each protocol we can associate a tuple of values signifying the value of these plausible future paths.

Definition 17. Given an arbitrary, fixed enumeration of $B_{cur}(l)$, the *plausible outcome* of a protocol $\mathcal{P}$ in l is a tuple whose kth element is the value of the run suffix generated by applying $\mathcal{P}$ starting at the kth state of $B_{cur}(l)$.

Notice how, in our example, the set of plausible worlds, which originally contained four members has been reduced to one following new observations. In general, when the agent acquires new information, its beliefs will change. In the current example, it was pretty much obvious how the robot's beliefs should change: the new information was consistent with the agent's previous beliefs, and it could be incorporated using a process much like probabilistic conditioning or logical conjunction. However, suppose that after its second motion command (another "go down"), the robot's position reading is $(3,1)$. This reading is inconsistent with its previous beliefs, under which only a reading of $(3,2)$ would be plausible, and simple conditioning yields an empty set of beliefs. In this situation, there are various ways in which the robot could revise its beliefs. For example, it could assume that current readings take precedence over future readings, in which case it will come to believe that its initial position was $(1,1)$ and that its first sensor reading was inaccurate. This issue of *belief change* is dealt with in Section 3.2. But first, let us consider preferences.

Unlike beliefs, we envision preferences to be quite stable. For example, suppose that our agent believes its position is $(3,6)$, and it is comparing two possible paths p and p' in both of which its next position would be position 2. If it prefers p to p' and its next reading is consistent with its belief that it has now reached position 2, it should prefer the remainder of p over the remainder of p'. More generally, suppose that the agent is uncertain about the state of the world, and it considers three initial states plausible: s_1, s_2, s_3. It has two possible protocols, one leading to the runs r_1, r_2, r_3 in each of the worlds, respectively; and one leading to the runs r'_1, r'_2, r'_3 in each of the worlds, respectively; and it prefers the first protocol. Now, suppose that both protocols assign the same action, a, to the agent's current local state and that after performing a, the agent's beliefs do not change, i.e., s_1, s_2, s_3 are still its plausible initial states. It is most natural to expect that the agent will still prefer the first protocol over the second protocol. Similarly, if we learned only now that the agent preferred the first protocol to the second, we would expect that it had similar preferences before performing a, since these protocols do not differ on the first step. Consequently, we assume that the following property, motivated by the intuition above, is satisfied by u and ρ.

Definition 18. Let $s \cdot r$, where $s \in \mathcal{E} \times L$ and $r \in \mathcal{R}_{suf}$, denote a run suffix whose first state is s, followed by the run suffix r. A decision criterion ρ is *static* with respect to a value function u if for any natural number k, for any set of run suffixes $r_1, \ldots, r_k, r'_1, \ldots, r'_k, s_1 \cdot r_1, \ldots, s_k \cdot r_k, s_1 \cdot r'_1, \ldots, s_k \cdot r'_k \in \mathcal{R}_{suf}$, we have that $(u(s_1 \cdot r_1), \ldots, u(s_k \cdot r_k))$ is at least as preferred as $(u(s_1 \cdot r'_1), \ldots, u(s_k \cdot r'_k))$ iff $(u(r_1), \ldots, u(r_k))$ is at least as preferred as $(u(r'_1), \ldots, u(r'_k))$.

That is, if we compare two tuples of values of run suffixes that have an identical tuple of states as their prefix, then the first tuple is more desirable than the second tuple if and only if the first tuple of values of the same run suffixes, but with the first state truncated, is more desirable than the corresponding second tuple.

Notice that given the fact that ρ is static with respect to u, it is easy to modify the definition of plausible outcome to use B_I instead of B_{cur}.

We have explained how the concepts of state, beliefs, and preferences change when we move to the dynamic model. In many respects, the decision process itself, remains as it was before. At each state, the agent compares a set of plausible outcomes and uses its value function and decision criterion (which we assume to be fixed) to choose the most desirable one. Consequently, the model ascription process is similar, and the constraint we must satisfy is that at each local state, the agent's actual choice conforms with the choice predicted by the model.

The main difference between the decision process in static and dynamic settings is that in order to match the notion of preference over runs, we must define the plausible outcomes based on sets of run suffixes rather than based on sets of states. This implies that we must view the agent as choosing between protocols rather than single actions, since a single action produces a single next state while a protocol produces a sequence of states. Indeed, our mobile robot must choose among different paths rather than different next positions; next positions cannot be judged "good" or "bad" by themselves but only in the context of the actions that follow. For example, the move from position 1 to

position (3,3) in Fig. 4 is good if it is followed by actions that lead to the displayed path. It is not good if it is followed by actions that lead the robot back to (1,1).

This view of the decision making process raises two practical issues: The first issue is why should the agent repeat the above comparison of protocols at each state; it can simply make the choice once and for all. Indeed, we will show that under certain natural conditions, a model in which the agent makes its choice once and for all is equivalent to the model described above. However, this is true only when the agent's beliefs change in a particular fashion and when its preferences have certain properties. We view this result as an indirect justification for adopting these constraints on the model. The second issue has to do with the feasibility of the above decision process. The size of the set of protocols is very large (exponential in the number of steps in the worst case). Hence, comparing all protocols is not a feasible alternative for the decision maker or for its modeler.[15] One solution would be to compare a single action at a time. Assuming a small, fixed set of actions, this is considerably simpler. However, this would require assigning values to single states, rather than the more immediately plausible approach of assigning values to run suffixes. In Section 3.3, we show when it is possible to use such a *decomposed* model in which the decision making process is reduced to choices between single actions in the current state.

3.2. Belief change

As the agent makes new observations, its local state changes and with it its knowledge. We can illustrate this process using Fig. 5. The agent is initially in local state l, where the possible initial worlds are v, w, x, y. It has two actions it can take, a and a', each leading to one of two new local states: a leads into l_1 when the initial world is v or x and into l_2 otherwise. a' leads into l_3 when the initial world is x or y and into l_4 otherwise. We see that after performing either action, the agent's knowledge increases because it considers fewer initial worlds to be possible. This need not always be the case, since the agent may "forget". We also see that the agent's new local state is a function of its action and the actual world. For example, if it performs a and the actual initial world is x, its new local state is l_1. However, if the actual world is y, its new local state is l_2.

With the change in the agent's local state following new observations, its ascribed mental state should change as well. Our mental-level model has two components which are not state-dependent, the agent's preferences and decision criterion, and one component which is state-dependent, its beliefs. In the static model, we did not require any relationship to hold between the beliefs of the agent at different local states, and the process of belief ascription could have been done locally. However, now local states are more closely related to one another through temporal order. We wish to add global constraints on the agent's beliefs that reflect these relations between local states. That is, we would like to model an agent's belief change.

[15] While this point makes intuitive sense, the process of choosing among protocols can, in certain contexts be implemented in a very efficient manner that does not involve an explicit comparison of all alternatives. It is important to remember that we are only committed to the semantics of this process.

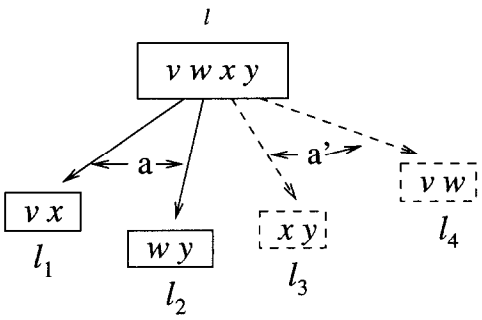


Fig. 5. The change in an agent's local state after performing actions a and a', respectively.

There is vast literature on the issue of how an agent should change its beliefs (e.g., [2,7,16,22,28,29]), and we will discuss its relation to our work later on. However, our modeling perspective will lead us to ask somewhat different questions.

In what follows, we assume our agent does not forget. Formally, an agent has *perfect recall* if its local state encodes all previous local states. This implies that the agent's local state is never the same in two different global states on a given run. Some of the following results also apply when this property holds under weaker conditions, e.g., the agent has a clock.

Consider the following restriction on belief change: my new plausible worlds should be those worlds that were previously plausible and are consistent with my new information whenever such worlds exist.

Definition 19. A belief assignment B_I is *admissible*,[16] if for local states l, l' such that l' follows l on some run, $PW_I(l') \cap B_I(l) \neq \emptyset$ implies that $B_I(l') = PW_I(l') \cap B_I(l)$. If $PW_I(l') \cap B_I(l) = \emptyset$, l' is called a *revision state* and the agent's new beliefs $B_I(l')$ can be any subset of $PW_I(l')$.

This manner of revising beliefs given consistent information can be viewed as the qualitative analogue of probabilistic conditioning, or alternatively to conjoining the new information to the agent's current beliefs. Indeed, this operation is quite standard in the literature on belief revision, which concentrates on restricting the agent's new beliefs in what we call revision states.

We can illustrate the process of belief change using Fig. 5. Assume that $B_I(l) = \{x, w\}$. After performing action a the agent finds itself in state l_1. If the agent's beliefs are admissible then $B_I(l_1) = \{x\}$. However, assume that $B_I(l) = \{x, v\}$ and the agent arrives at l_2 after performing a. Now we cannot say anything about the agent's beliefs at l_2, even if its beliefs are admissible (except of course $B_I(l_2) \subseteq \{w, y\}$). Clearly, the agent's plausible worlds in the past are not really possible, and it must now revise its beliefs to reflect this.

[16] This is unrelated to the game-theoretic notion of admissibility.

If we were to assume that the set of possible worlds consists of models of some theory then, in syntactic terms, admissibility corresponds to conjoining the new data with the existing beliefs, whenever this is consistent.

We can understand this restriction better through the following representation theorem. This theorem shows that we can either ascribe the agent beliefs that change locally in accordance to the admissibility requirement or we can ascribe the agent a more complex static ranking structure that uniquely determines its beliefs in each state. Specifically, at each state l the set $B_I(l)$ is exactly the set of elements in $PW_I(l)$ that are minimal with respect to this ranking.

Definition 20. A *well founded ranking* r of a set Q is a mapping from Q to a well ordered set $\mathcal{O}$ (which we will take to be the integers). Given a subset Q' of Q, the elements minimal in Q' are those that are assigned the minimal rank among the ranks assigned to elements of Q'.

A ranking of Q associates each member of Q with the group of other members having the same rank and orders these groups according to the rank assigned to them. In general one speaks of a total pre-order with minimal elements. The elements of lower ranks are considered better, more preferred, or more likely.

Theorem 21. *Assuming perfect recall, a belief assignment B is admissible iff there is a ranking function r (i.e., a total pre-order) on the possible initial worlds such that* $B_I(l) = \{s \in PW_I(l) \mid s \text{ is } r\text{-minimal in } PW_I(l)\}$.

Patterns of belief change similar to ours emerge in the work of other researchers (e.g., [22,36]). Indeed, relations between belief revision and belief update, and representations using partial and total pre-orders are well known. It was shown by Grove [25] that any revision operator that satisfies the AGM postulates [2] can be represented using a ranking of the set of possible states. However, to obtain that result, additional constraints on the agent's beliefs in a revision state are needed. We do not require such constrains because we are looking at a special case of belief revision: our agent has perfect recall and we do not need to account for general counterfactual reasoning. We learn about the agent's response to counterfactual queries when it receives information that contradicts its beliefs; we called this situation a revision state. However, when one observes an agent acting in the world, only a limited number of such revisions can occur, all of which must be consistent with the actual state of the world. This puts less constraints on the modeler and allows us to obtain this result. When we observe an agent repeatedly performing the same task, starting at different actual worlds but with the same local state, we will need the additional properties used in [25,29] to obtain a fixed ranking. However, we note that the difference between our approach and the AGM approach [2] is more fundamental. They ask the question: how should I change my beliefs? We ask: how should I model the belief change of another agent? This difference becomes clearer in the next subsection, where we examine the suitability of admissible beliefs for modeling agents.

Finally, in [10] we investigate another pattern of belief change, which we call *weak admissibility*. Weakly admissible beliefs allow the new plausible worlds to contain possible worlds that were not plausible before, even in non-revision states.

3.3. Ascribing admissible beliefs

Having defined mental-level models for dynamic agents, we come to the question of their construction. The general process of ascribing a mental-level model and the special case of ascribing beliefs remain the same with the transition from static to dynamic agents. Much like the static case, we must search for models that are consistent with the agent's behavior (i.e., its protocol).

Definition 22. A mental-level model for a dynamic agent $\langle \mathcal{L}_{\mathcal{A}}, A_{\mathcal{A}}, B_l, u, \rho \rangle$ is *consistent* with a protocol $\mathcal{P}$ if for all $l \in L$ it is the case that the plausible outcome of $\mathcal{P}$ is most preferred according to B_l, u, and ρ, and B_l is admissible. It is consistent with a partial protocol $\mathcal{P}'$ if it is consistent with some completion $\mathcal{P}$ of $\mathcal{P}'$.

Two properties of dynamic agents seem to make ascribing a mental-level model more difficult. Both of these properties have to do with an apparent loss of locality, in terms of what beliefs are acceptable and in terms of how we evaluate the effect of actions. The first problem is that in the dynamic case an agent's beliefs in one state are constrained by its beliefs in other states. Thus, we cannot decompose the ascription of beliefs. The second problem is that the plausible outcome of a protocol $\mathcal{P}$ given $B(l)$ is no longer dependent only on the action that $\mathcal{P}$ assigns to l. That is, we no longer measure the effect of actions locally because we use a value function over runs rather than states. Therefore, we cannot say how good a single action is in isolation from the actions that follow it, and we must compare complete protocols instead of single actions. For example, buying a run-down apartment may lead to a good outcome if I plan to renovate it later and sell it at a premium, or it may lead to a bad outcome if I use it as my place of residence. As we will see, under the following weak condition on the agent's decision criterion, both of these problems can be handled. Here, we use $\circ$ to denote vector concatenation.

Definition 23. Let (w, w') and (v, v') be two pairs of real valued vectors such that $|w| = |v|$ and $|w'| = |v'|$. A decision criterion satisfies the *sure-thing* principle if $v \circ w$ is at least as preferred as $v' \circ w'$ whenever v is at least as preferred as v' and w is at least as preferred as w'.

Intuitively the sure-thing principle [56] says the following: suppose you prefer action a over a' when the current (or initial) plausible worlds are w_1, w_2, w_3, and you also prefer a over a' when the plausible worlds are w_4, w_5, w_6. Then, you should prefer a over a' when the plausible worlds are $w_1, w_2, w_3, w_4, w_5, w_6$.[17]

[17] Our definition of the sure-thing principle is not the same as Savage's original definition because of differences in the basic framework. However, the essential idea is the same.

Throughout this section, we restrict ourselves to mental-level models in which the decision criterion satisfies the sure-thing principle. Under this assumption, we can show that the constraints imposed on beliefs at local states by the admissibility requirements make the ascription process easier.

Theorem 24. *Let $\mathcal{P}$ be an agent's protocol. If this agent 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.*

That is, if we are able to find a consistent assignment of belief to an agent at its initial state based on a given protocol, we are guaranteed that in the following (non-revision) local states the belief assignment that is obtained by following the criterion of admissibility is also consistent. Revision states are not constrained by the initial belief assignment, and ascribing beliefs in these states is analogous to the task of ascribing beliefs at initial states. Hence, admissibility, rather than being a handicap is actually an advantage.

The second problem stemmed from the fact that in the dynamic case, the natural definition of values is over runs. If we could provide conditions under which a natural definition of values over states is possible, we would not have this problem. This will allow us to construct a simpler, decomposed model with which it is easier to work. In this model, we will need to consider the immediate effect of actions only, rather than the long-term effect of protocols.

Definition 25. A *local value function* is a function u_{cur} from the set S of global states to $\mathbb{R}$. Given a fixed enumeration of the elements of $B_{cur}(l)$, the *plausible local outcome* of an action a at l is the (suitably ordered) tuple containing $u_{cur}(a(s))$ for each $s \in B_{cur}(l)$.

With these ideas we can proceed to define a decomposed mental-level model in which both the beliefs and the values are localized. That is, the value function is defined over states, rather than run suffixes. Consequently, verifying that a decomposed model is consistent with a protocol $\mathcal{P}$ (or our observations) is much easier: For each local state l, we use the state based value function to check whether the action $\mathcal{P}(l)$ that is assigned by $\mathcal{P}$ at l has the most preferred immediate outcome.

Definition 26. A *decomposed* mental-level model is a tuple $\langle L, actions, B_{cur}, u_{cur}, \rho \rangle$.

A decomposed mental-level model $\langle L, actions, B_{cur}, u_{cur}, \rho \rangle$ is *consistent* with a (possibly partial) protocol $\mathcal{P}$ if

(1) for any local state l on which $\mathcal{P}$ is defined, the plausible local outcome of $\mathcal{P}(l)$ in l is most preferred among all plausible local outcomes;

(2) if l' follows l on some run, then

$$B_{cur}(l') = PW_{cur}(l') \cap \{\mathcal{P}(l)(s) \mid s \in B_{cur}(l)\},$$

when it is not empty.

(This last condition says that B_{cur} is admissible.)

Given a local value function over states, in order to ascribe beliefs to the agent, we no longer have to examine the effects of complete protocols and compare the values of run suffixes. Instead, we simply compare the immediate plausible outcome of single actions, much like the case of static agents. Hence, decomposed models are conceptually simple, are simpler to compute with, and require a simpler representation.

In the following we will identify runs of bounded length with runs containing a suffix all of whose states are identical. We can show the following:

Theorem 27. *Let $\mathcal{A}$ be an agent whose possible runs have bounded length, for which we can ascribe a consistent mental-level model $\langle L, actions, B_I, u, \rho\rangle$, where B_I is admissible and ρ satisfies the sure-thing principle. Then, $\mathcal{A}$ can be ascribed a consistent decomposed mental-level model $\langle L, Actions, B_{cur}, u_{cur}, \rho\rangle$.*

A corollary of these results is thus:

Corollary 28. *Let $\mathcal{P}$ be the 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 under a decision criterion that satisfies the sure-thing principle. Then, this agent can be ascribed an admissible belief assignment at all local states and a local value function* over states *such that its observed* action *has a most preferred plausible outcome according to the local value function.*

In practice, replacing a standard value function (over runs) with an equivalent local value function is quite difficult, requiring a process analogous to dynamic programming. Consider an agent playing chess against a computer program. A value function can be assigned naturally to a run suffix based on whether or not the agent wins in this run. A local value function would tell the agent at each state whether a particular move will lead it to a position from which it can win, tie, or lose. Hence, a local value function is really an ideal heuristic. Naturally, computing such a function is often unrealistic, but so is the task of comparing all possible protocols. In practice, a compromise is struck using some form of look-ahead and some local evaluation function, i.e., a less than perfect heuristic function.

4. Prediction

One central application of agent modeling that is of particular interest in the multi-agent context is prediction. In this section we pause to show how the various aspects of our theory can be combined to obtain an approach for action prediction using a mental-level model. The context we would like to deal with is the following: We observed an agent taking part in some activity; we know its goals; and we wish to predict its next actions. The approach we suggest underlies some of the work done on plan recognition and some of its applications to air-combat simulation, discussed in Section 9.

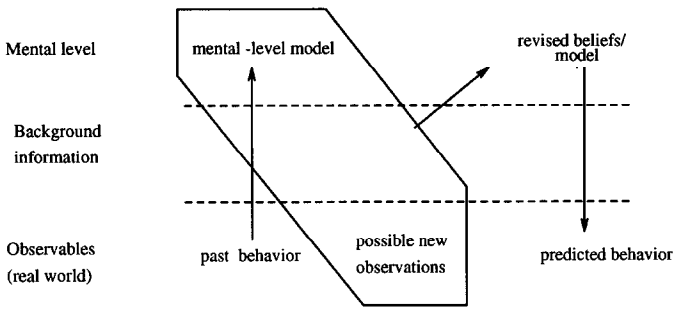


Fig. 6. Three step prediction process.

To predict an agent's next action, we go through three steps (illustrated in Fig. 6):

(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; and
(3) predict the action (or protocol) which has the most preferred perceived outcome based on these beliefs.

The following example serves to illustrate this idea.

Example 29. We start with a robot located at an initial position whose task is to find a small can located in one of three possible positions: A, B, or C. We assume that the robot knows these to be the only possible positions of the can. Hence, we have three possible initial states of the environment. The robot can move in any direction and can recognize the can from a distance of 2 meters. (See Fig. 7.)

In this example, a run would describe 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.

We will assume that the following value function (over runs) describes the robot's preferences, which depend on the length x of the robot's trajectory, and on whether or not it terminates in the position of the can,

$$u \stackrel{\text{def}}{=} 10 - x + 20 * \begin{cases} 1 & \text{if the trajectory terminates at the can,} \\ 0 & \text{otherwise.} \end{cases}$$

Having observed the robot's initial path, as shown in Fig. 8, we can try to ascribe it beliefs using our background knowledge of its preferences and the assumption that it uses the maximin criterion (although the following also applies to the principle of indifference). What we see is that the ascribed plausible initial worlds are those in which the can is in $\{A, B\}$ or $\{A, B, C\}$. That is, only with such initial plausible worlds would the robot choose the observed trajectory. For example, if the robot believed only one initial possible world to be plausible it would head directly to the can's position in that state. Similarly, if $\{B, C\}$ was believed, the robot's path would head more toward them.

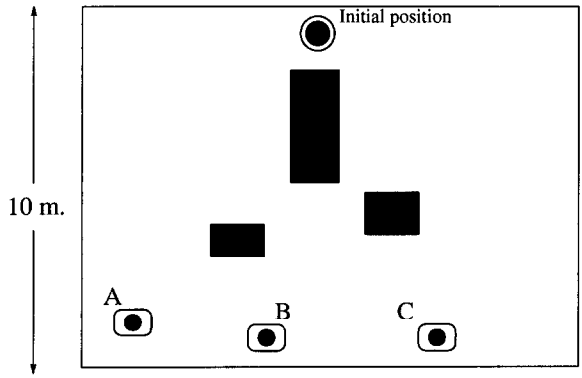


Fig. 7. Initial set-up.

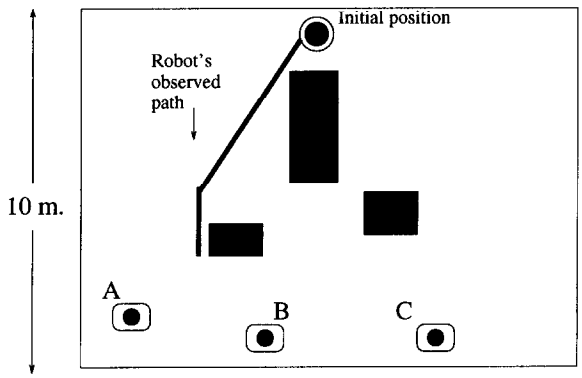


Fig. 8. Robot's initial path.

Next, suppose that the can is in B. At its current position the robot can see that the can is not in A, and its local state has changed to incorporate this knowledge. The robot's new beliefs are now revised to contain B and possibly C (assuming its beliefs are admissible). Given these beliefs, we can predict that the robot's next action would be to turn to its left (i.e. toward B).

One weakness of this approach is that we have little to say when the modeled agent's observation contradicts its beliefs (what we called revision states). Recall that our definition of admissibility in Section 3.2 does not constrain the agent's new beliefs under these conditions. Recalling the relationship between an admissible belief assignment and a ranking over the set of possible worlds discussed there, we can see this problem could be overcome if we have observed this agent in the past and have learned its ranking. Then, we can use this ranking to deduce the agent's new beliefs even in a revision state. Another approach is possible when our set of possible worlds has more structure to it, in which case we can attempt to induce the ranking, or some of its properties. Indeed,

most often we equate the set of possible worlds with a set of models of some language, in which case different worlds are related by the sentences to which they assign similar truth values. We can then introduce various assumptions about the relationship between the agent's beliefs, as stated in that language, before and after new observations are made. Various relationships appear in the literature, and a number of such methods are discussed in [29]. As an illustration, consider an agent that knows $p \vee r$ and believes $p \wedge q$. Suppose it now learn $\neg q$. Thus, none of the worlds that were previously considered plausible are still plausible. However, we could still assume that its belief in p persists.

5. Choosing among belief assignments

As we observed in the thermostat example, often there is more than one consistent belief assignment. Indeed, we find this to be the case even in simple examples. We can handle this problem by using background knowledge to restrict further the kind of models we are willing to assign. One common approach for choosing among different models is to a priori restrict or rank them. In the first case, we limit the models that we are willing to ascribe; this is similar to the *restricted hypothesis space bias* in machine learning [17]. In the latter case we use the ranking over models to ascribe the most normal consistent model; this is similar to the notion of *preference bias* [17].[18] In particular, we could use the additional structure obtained when the set of possible worlds corresponds to models of some propositional language. Then, we restrict the class of models we are willing to ascribe to those in which the belief assignments correspond to relatively simple formulas, such as conjunctions of primitive propositions.[19] Alternatively, we can use some measure of the complexity of the formula to rank the different models.

In this section, we investigate two domain-independent choice heuristics and show that under certain conditions they lead to a unique choice of model. We start with some general heuristics for the static model and continue with a particular heuristic for the dynamic model.

5.1. Choice assumptions for static agents

A common modeling bias is to favor models that offer adequate explanation of the data. In our case, we would like the ascribed model to be such that, at each state, there is a unique most preferred plausible outcome rather than a set of most preferred plausible outcomes.

Definition 30. A mental-level model $\langle \mathcal{L}_{\mathcal{A}}, A_{\mathcal{A}}, B, u, \rho \rangle$ is *explanatory* if for all local states l, the decision criterion ρ returns a unique plausible outcome when applied to the set of plausible outcomes of all protocols in l.

[18] In practice, it seems that people impose biases that make other people's beliefs or preferences similar to their own.

[19] This was pointed out to us by Hector Levesque.

Example 2 (*continued*). Recall that in Section 2.3 we were able to constrain the thermostat's beliefs in the state "−" to only those that include the state *cold*. Four belief assignments satisfy this property. However, only one of them, $B(-) = \{cold\}$ is explanatory. Given this belief assignment the agent *must* choose the action *turn-on*, while given any of the other three belief assignments, the agent is indifferent to the choice between *turn-on* and *shut-off*.

A different modeling bias is toward greater generality. Given a number of possible models that explain some behavior *equally well*, the preference is for those making fewer assumptions regarding the agent's beliefs. That is, we prefer belief assignments in which fewer worlds are ruled out.

Definition 31. A belief assignment B is *more general* than B' if for all $l \in L_{\mathcal{A}}$ we have that $B'(l) \subseteq B(l)$ and $B \neq B'$.

Given a set of belief assignments, $\mathcal{B}$, $B \in \mathcal{B}$ is a *minimal* belief assignment with respect to $\mathcal{B}$ if there is no $B' \in \mathcal{B}$ such that B' is more general than B.

Thus, minimal belief assignments ascribe to the agent the weakest set of beliefs. That is, they exclude as implausible the smallest number of possible worlds. A belief assignment is *minimal explanatory* if it is a minimal belief assignment among those belief assignments for which the mental-level model is explanatory. That is, a minimal explanatory belief assignment rules out just enough worlds to be explanatory.

Example 2 (*continued*). Any belief assignment that is a non-empty subset of $\{ok, hot\}$ is explanatory for the state +. However, there is a unique minimal explanatory belief assignment for that state: $\{ok, hot\}$.

To summarize, we have the following (unique) minimal explanatory belief assignment for the thermostat:

state	−	+
belief	cold	not-cold

At this stage we believe we have a satisfactory formal account of McCarthy's thermostat example, which has been useful in demonstrating our basic concepts. In this example there was a unique minimal explanatory belief assignment. We now proceed to examine whether this is true in the general case.

Example 32. Consider the following decision problem:

	s_1	s_2	s_3	s_4	s_5
a_1	2	2	11	2	2
a_2	7	7	0	7	7

Suppose the agent has taken action a_1. We can see that both $\{s_1, s_2, s_3\}$ and $\{s_3, s_4, s_5\}$ are consistent (explanatory) belief assignments given the principle of insufficient reason as a decision criterion. However, there is no unique minimal consistent belief assignment because it would have to contain both of these belief assignments, hence it would have to be $\{s_1, s_2, s_3, s_4, s_5\}$. However under this belief assignment, the action a_2 is more preferred.

We just saw that a unique minimal belief assignment does not always exist. However, for belief assignment problems with certain decision criteria, a unique minimal belief assignment and a unique minimal explanatory belief assignment exist.

Definition 33. Let (v, v'), (x, x'), and (w, w') be three pairs of equal length vectors of reals. A decision criterion is *closed under unions* if $v \circ w \circ x$ is at least as preferred as $v' \circ w' \circ x'$ whenever $v \circ w$ is at least as preferred as $v' \circ w'$ and $w \circ x$ is at least as preferred as $w' \circ x'$.

When we substitute the empty vector for w and w' in this definition, we obtain the sure-thing principle. Thus, a decision criterion that is closed under unions satisfies this principle. An intuitive reading of this property is the following: suppose that ρ prefers action a over a' when the plausible worlds are x, y, z and it also prefers a over a' when the plausible worlds are v, w, x. If ρ is closed under unions, it would prefer a over a' also when the plausible worlds are v, w, x, y, z.

The principle of insufficient reason is not closed under unions, hence the lack of unique minimal belief assignment in Example 32. However, the maximin criterion is closed under unions, as are a number of other criteria discussed in [41].

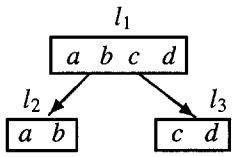
Theorem 34. *Given a belief ascription problem with a decision criterion that is closed under unions, if a consistent belief assignment exists then there is a unique minimal consistent belief assignment.*

In addition, the above theorem holds when we replace *consistent* with *explanatory and consistent.*

5.2. *Choosing among admissible beliefs*

In modeling dynamic agents, we prefer to use admissible belief assignments in which the agent's beliefs change coherently over time. However, for such beliefs, the minimality bias makes little sense without some additional modifications. As we now show, it is possible to have less general beliefs at some times while having more general beliefs at other points.

Example 35. Consider the following case: there are four possible initial worlds in the local state l_1: a, b, c, d. After the first action there are two possible local states l_2 and l_3 with worlds a, b possible in l_2, and worlds c, d possible in l_3.



Assume that in l_1 we can consistently ascribe beliefs in $\{a, b\}$ or $\{a, b, c\}$, in l_2 we can consistently ascribe beliefs in $\{a, b\}$, and in l_3 we can consistently ascribe belief in $\{c\}$ or $\{c, d\}$. Therefore, there are two consistent admissible belief assignments: B_1 assigns $\{a, b, c\}$ to l_1, $\{a, b\}$ to l_2 and $\{c\}$ to l_3 while B_2 assigns $\{a, b\}$ to l_1, $\{a, b\}$ to l_2 and $\{c, d\}$ to l_3, none of which is more general than the other. Note that $B_1 \cup B_2$ is not admissible.

Hence, in the case of dynamic agents with admissible beliefs, we will have to come up with a weaker notion of generality. What seems to us most appropriate is to prefer generality at earlier points of time. That is, we prefer to model the agent as setting out with a minimal initial belief assignment, making as few initial assumptions about the world. This also implies that it will have fewer revision states. In addition, in revision states we would also prefer models in which the agent makes fewer assumptions. Described in terms of the static ranking associated with agents whose beliefs are admissible, this type of preference over models translates into a preference of rankings that are "thicker" at the bottom. Indeed, in the context of non-monotonic logics such preferences are well known. In non-monotonic logics, worlds minimal in a ranking structure are often described as "most normal", and structures that are "thicker" at the bottom are often preferred because they make fewer assumptions of non-normality (see, e.g., [37]).

Definition 36. An admissible belief assignment B_I is *more general* than B'_I if, represented as ranking functions, B_I and B'_I are identical up to some rank m, and $B_I^m \supset B_I'^m$ (where B_I^m is the set of states in the mth rank of B_I).

We will refer to this definition of more general in the context of admissible beliefs and in the resulting definition of minimal belief assignments.

Theorem 37. *For agents with perfect recall, if the decision criterion is closed under unions then the set of consistent admissible belief assignments, if non-empty, contains a unique minimal belief assignment.*

Again, the above theorem holds when we replace *consistent* with *explanatory and consistent.*

6. Existence

Given any approach to modeling agents (or other entities for that matter), one of the most important questions is under which conditions this model is adequate. That

is, what kind of behaviors can we model using this approach. Or alternatively, if we decide to adopt the model, what implicit assumptions are we making about the agent's behavior. We are finally in a position to provide a partial answer to this question.[20]

We approach this task in a manner similar to Savage's work on the foundations of subjective probability [56]. Savage's approach allows us to ascribe a probability assignment and a value function to an agent based on its choice among actions. Similarly, we will describe a class of situations and a number of restrictions on the agent's behavior under which a mental-level model with a unique minimal admissible belief assignment and a unique minimal explanatory admissible belief assignment can be ascribed.

First, notice that we can trivially ascribe a consistent mental-level model to any agent. We simply make all runs have the same value and choose any decision criterion and any beliefs. This observation shows that we should ask more constrained questions, such as, under what conditions can we solve a belief ascription problem or when can we ascribe an explanatory model. We will now characterize a class of agents for which beliefs can be ascribed

Goal-seeking agents are agents with perfect recall whose local states are of two types: goal states and non-goal states. These agents have a special action, called HALT, which intuitively signals the end of a run and must eventually be performed at each run. The value of a run is determined by the state in which HALT is performed: it is 1 if HALT is performed in a goal state and 0 otherwise. Goal-seeking agents are quite natural from the AI perspective, since they describe agents that act to bring about a particular state of the world.

The protocols of goal-seeking agents satisfy two *weak* rationality postulates that embody a minimal notion of goal-seeking behavior. The *rational effort* postulate says that the agent must halt whenever it is in a goal state or when it is impossible to reach a goal state. Thus, the agent does not perform actions unless they can somehow help it attain a good state—its efforts are rational. The *rational despair* postulate says that to halt the agent must either be in the goal, or it must be able to show a possible world under which it can never reach the goal. Thus, the agent does not give up without reason, and its despair is rational. (A stronger postulate would require it to stop acting only when it is impossible to reach that goal, no matter how unlikely the prospect of reaching the goal is.) Notice that these postulates refer to the set $PW_{cur}(l)$ rather than to $B_{cur}(l)$ (preventing possible circularity later).

Rational Effort Postulate. The protocol either assigns HALT to a local state l, or it assigns an action that weakly dominates HALT.[21]

Hence, unless there is something better to do, HALT is assigned.

[20] Additional results of this nature appear in [12,13].

[21] Given a tuple v of length k, let $v(i)$ be its ith element. We say that v' *weakly dominates* v if for every $1 \leqslant i \leqslant k$, it is the case that $v'(i) \geqslant v(i)$ and for some $1 \leqslant i \leqslant k$, $v'(i) > v(i)$.

Rational Despair Postulate. The protocol in a non-goal local state l is HALT only if for some $s \in PW_{cur}(l)$ there is no protocol that achieves the goal.

Hence, HALT is assigned in a non-goal state only if it is possible that the goal is unachievable.

Finally, goal-seeking agents use weak dominance as their decision criterion, i.e., they strictly prefer w over v iff w weakly dominates v.

An example of a goal-seeking agent could be an ordinary mobile robot with a goal state, some motion command (telling it in which direction to move at each state), and some termination condition (telling it when to HALT). Typically, such robots are not programmed using a knowledge-based paradigm but rather, they employ a motion planner or an ordinary planner. In that case, the two postulates translate into two weak requirements on the robot's termination condition. For instance, a robot that stops only when it is in the goal or if it is in a component of its configuration space that is not connected to the component containing the goal, will satisfy the two postulates. The condition that it employs weak dominance as its decision criterion would be consistent with most reasonable motion strategies. In that case, as the following theorem shows, we can ascribe beliefs to such a robot.

Theorem 38. *If A is a goal-seeking agent then it can be ascribed a unique minimal admissible belief assignment and a unique minimal explanatory admissible belief assignment.*

If we allow a decision criterion that is stronger than weak dominance (but consistent with it), we may have to drop uniqueness from the statement of this theorem.

We can use this result to show that agents with perfect recall and a HALT action can be ascribed a mental-level model with a fully explanatory admissible belief assignment, i.e., in which the actual protocol is strictly preferred over all other protocols. We need to show that we can attach to these agents goals so that they satisfy the criteria of goal-seeking agents. We do so by ascribing to the agent a value function in which all runs that can be obtained using its protocol have value 1, while all other runs have value 0.

Many people view rational choice as equivalent to expected utility maximization under some probability distribution. We show that in 0/1 value contexts any behavior consistent with expected utility maximization under some probability distribution can be attributed belief in our framework. Let us define a *B-type* agent as one whose beliefs are represented by a probability assignment, whose preferences are represented by a 0/1 value function, and whose decision criterion is based on the maximum expected utility principle. We require only that the agent perform HALT when no action has an expected value greater than 0.

Corollary 39. *If an agent can be modeled as a B-type agent, it can also be modeled as a plausible outcome maximizer that uses some admissible belief assignment and a decision criterion consistent with weak dominance.*

7. Related work in economics and game theory

There is much work within AI that is relevant to the framework presented in this paper, and this work will be discussed in the next section. However, the most closely related lines of research can be found within economics, game theory, and decision-theory, areas whose relevance to AI research has been pointed out by many researchers, most notably, Jon Doyle (e.g., [18, 19]). In particular, three topics of research are directly relevant to our work: work on subjective probability and choice theory, work on qualitative decision making, and work on revealed preference. In what follows, we briefly describe these fields and compare them with our own effort.

7.1. Subjective probability and choice theory

A key issue for economists and game-theoreticians in building models of economic systems is how to model agent behavior. An agent model must be descriptively accurate, i.e., it must correctly predict human behavior in economic contexts. Moreover, the model should be amenable to mathematical analysis, so that it can be used in practice. One of the most popular approaches has been to model economic agents as entities with a mental state consisting of beliefs, preferences, and a decision criterion, much like our model. However, rather than use the qualitative models described in this paper, beliefs are modeled using a probability distribution, preferences are modeled using a utility function, and the decision criterion is expected utility maximization. While this model does not necessarily offer a computational advantage, it is mathematically appealing, partly because continuous mathematics offers powerful analysis tools and because probability theory is well developed. More importantly, it allows us to represent finer degrees of belief and preference as well as risk attitudes.

But while the quantitative probabilistic model is quite elegant, it is by no means clear that it is an adequate model of human behavior. In particular, most people do not feel they perform expected utility calculations when making different choices. But here lies an important conceptual idea upon which the theory of modeling choice [32] is founded: whether or not the agent actually makes its decision using probabilistic reasoning is of no consequence. The issue is whether or not a probabilistic model employed externally has the required predictive power. That is, will this model lead to accurate predictions of *observables*. This idea is central to our modeling approach as well, and it is the basis of Newell's concept of the knowledge level.

These considerations lead to the emphasis placed within game theory on existence, or representation theorems. These theorems tell the modeler under what assumptions on the agent's behavior the model is applicable. Because the assumptions are on the agent's behavior which is, in principle, observable, they can be empirically tested and (in)validated.

The work of Savage [56] provides what many consider to be the most important result in choice representation. Savage provides a set of assumptions about the agent's approach to action choice under which it can be modeled *as if* it were an expected utility maximizer. Again, this does not mean that the agent performs expected utility

calculations in his/her head, but that such a model would make correct predictions of his/her behavior. One of Savage's famous conditions is the *sure-thing principle.*

Many people find Savage's assumptions to be quite intuitive from a normative point of view. However, their descriptive adequacy is not clear and considerable effort has been expanded by psychologists in testing their validity (see e.g., [42]). However, the relevance of these studies to the descriptive adequacy of probabilistic modeling of artificial agents is not clear. In order to obtain better descriptive models, various weaker representation theorems have been proposed. These theorems make weaker assumptions on the manner in which agents choose their actions and use weaker representations of beliefs in their agents models, e.g., non-additive probabilities (see [21] for more details).

Savage's seminal result, as well as many following works, make two strong requirements that render their application in our setting quite difficult: One must supply a total pre-order on *all* functions from initial states to outcomes, many of which correspond to no existing real-world action; this information will not be available to an observer of the system. Moreover, they require a rich state description, where for any natural number n, there exists a partition of the set of states into n subsets, all of which are equally likely. Thus, the number of states must be infinite.

It is still early to compare the well developed theory of choice with our approach, and here we simply note a number of differences. Our formalism deals with discrete descriptions of mental state. Therefore, it can provide semantic foundations for the use of these mental states in agent models. This is an important consideration given the abundance of work on discrete notions of belief and the prevalence of qualitative representation tools within AI. Moreover, by varying the decision criterion we may be able to cover different classes of agents. For example, one type of decision criteria we are currently looking into takes into account the agent's limited computational resources. Such criteria may be better suited for modeling actual agents. Finally, our approach does require substantial background information to be applicable. However, because it is discrete and qualitative, it should require less information than that needed to specify a quantitative probabilistic model. In addition, we hope that the heuristics for model choice, suggested in Section 5, can lessen this burden.

7.2. *Qualitative decision theory*

While Savage's work is of great importance to game-theorists, it is of no lesser value to decision theorists concerned with the question: how *should* one make one's own decisions. While there are doubts as to the descriptive adequacy of Savage's postulates, there is much less disagreement about their normative appeal.[22] Because of this appeal and consequent progress in the disciplines of decision analysis and Bayesian statistics, the last few decades have seen very little work in qualitative decision theory.

Work in qualitative decision theory has mostly been concerned with the following question: how should one make decisions when one has virtually no information about the likelihood of different states of the world but a good assessment of the desirability

[22] Though there is no consensus on this matter. See the articles in [23].

of different outcomes. This is often referred to as decision making under complete ignorance. All of the decision criteria that were mentioned in this paper, i.e., *maximin*, *minmax regret* and the *principle of indifference*, have been studied in this context [41]. However, until recently, there were no representation theorems analogous to Savage's for these qualitative decision models. In [12,13], we present the first such results for the *maximin* and *minmax regret* criteria. Using the language of this paper, these results should be viewed as existence theorems for static agents employing these decision criteria.

Our current work has been motivated by the models used in decision making under ignorance. However we modified these models to capture qualitative information. Rather than assume that all states of the world are possible, as in the above works on qualitative decision making, we incorporated a qualitative notion of belief that allows the agent to discriminate between more likely and less likely worlds.

7.3. Revealed preference

Besides the technical differences, there is an important, but subtle difference between choice theory and our work. Both approaches desire sound foundations for abstract models of agents. However, choice theorists wish to justify the use of probabilities and utilities in modeling the behavior of *generic* human agents in general economic theories. That is, they do not have the engineering motivation of monitoring a particular, actual person (or agent) with the goal of constructing a model of this particular person in order to explain his/her behavior and predict his/her future behavior.

Such concerns, which are central to our approach, motivate the work on revealed preference in economics. The goal of revealed preference theory is to predict future preferences of single agents and classes of agents. For example, based on previous consumption habits of an agent, we may be able to predict its future habits. Such information can be of considerable value to many economic agents. The basic idea is that an agent's choices in various settings reveal his/her preference among various options. For example, the consumer problem in economics is that of choosing a good bundle x (which can be thought of as a vector specifying the quantity of various goods) given a certain price vector p and available income y, such that x is the best choice under the constraint $x \cdot p = y$. If we observe a choice x made by a consumer under given y and p, we can deduce that this consumer prefers x over all other x' that satisfy the given constraint.

Under the assumptions that preferences remain stable we can combine a set of observations (x_i, y_i, p_i) together with basic assumptions on preferences to "reveal" the consumer's general preference ordering over bundles. For example, one property, called the *general revealed preference principle* stipulates that preferences are acyclic. We refer the reader to [33] for more details.

The aims of revealed preference theory and of our work have considerable overlap. In both cases, an attempt is made to construct an agent model with predictive power. In both areas, the question of preference persistence arises. The main difference is that revealed preference theory does not attempt to explain the observed preferences in terms of a more basic agent model. Thus, the types of predictions made by revealed preference

theory are somewhat like the following: First, an agent is observed passing an object from its left side, rather than its right side. Next, it is seen passing an object from the right, rather than going over it. Consequently, it is concluded that, when possible, the agent would prefer passing an object from the left side rather than passing it from above. Such deductions are important, but their scope is narrower.

8. Related work in AI

The understanding of mental states has steadily progressed through the effort of various researchers, and we have greatly benefited from many existing ideas. We proceed to discuss some of the more relevant work.

Structure

A large portion of the AI literature on formalizing mental states deals with distinct mental attributes such as belief and knowledge, and their dynamics. However, a number of researchers suggested more complete models of mental state, e.g.,[6, 49, 52, 55, 57]. While their aim has been to supply intuitive and well founded tools for agent specification and design, rather than agent modeling, they are clearly relevant to the question of what structure our model should have. Rao and Georgeff [52] define an interpreter that uses three mental components, beliefs, intentions and desires. Pollack et al. propose an abstract agent architecture based on similar mental states [49]. Rosenschein and Kaelbling [55] developed an interpreter that can implement behavior that is specified using notions such as knowledge and goals. Shoham [57] presents an agent oriented programming language based on the notions of belief and commitment. What these structures lack is the notion of a decision criterion, which embodies the agent's approach to action choice under uncertainty. While the need for deliberation under uncertainty has not escaped the attention of AI researchers (e.g., Thomason [60] incorporates some type of common-sense deliberation about conflicting goals, and Rao and Georgeff [51] incorporate expected payoff calculations for decision making), qualitative decision criteria have only recently shown up in the AI literature on mental states [6, 10].

In contrast, we do not include intentions in our model. It seems that intentions play an important role in modeling resource bounded agents: much like beliefs allow the agent to ignore certain possible outcomes of its actions, intentions allow it to ignore certain possible actions. Thus, it may be desirable to integrate intentions into future models. Nevertheless, the current model contains the three essential aspects of the mental state of any agent acting in the world: perceptions (beliefs), goals (values), and a method for choice under uncertainty (decision criterion).

Grounding

A number of important works ground single mental attributes, specifically, belief or knowledge. Methods of grounding these attributes are useful in our modeling context when they show us how to model a particular agent, and when they are able to say whether an agent is implementing a particular mental-level specification.

Halpern and Moses [26] and Rosenschein [54] ground the notion of knowledge in the relationship between the local state of a machine and the state of its environment. We discussed their work in Section 2. This research was the first to ground a mental state in a computer science context, and we were motivated by the desire to follow their lead, but with a more comprehensive description of an agent's state. A model of an agent's knowledge does not tell us about this agent's actual behavior.

Other works have proposed groundings for beliefs. We presented our view of the semantics of beliefs in the end of Section 2.3. Bacchus, Grove, Halpern, and Koller [4] ground the notion of belief in statistical information. Their work answers questions such as: what should we believe about the bird Tweety given that 90% of birds fly. Statistical information can be viewed as summarizing concrete observation of the world, therefore, it is grounded. However, one should notice that the question they pose is normative, not descriptive. While statistics can help us form beliefs about how thermostats act, they do not provide meaningful answers to the question of what beliefs we should ascribe the thermostat. Even in the case of an "intelligent" agent, using those ideas to ascribe its beliefs requires knowing what statistical information *it* has, and that *it* is acting in accordance to the ideas of Bacchus et al.

Goldszmidt and Pearl's ε-semantics [1,24] views beliefs as qualitative representations of probabilities: One believes that birds fly if one holds that $\Pr(Fly|Bird) \approx 1$. This approach is semantically close to ours, since subjective probability can be given semantics in terms of the agent's choice of actions. However, it would provide a roundabout manner of modeling agents: first, we ascribe them probabilistic beliefs and then, we discretize these beliefs. As we have noted, there are some difficulties in ascribing probabilistic beliefs to agents, and moreover, the discretization process used by ε-semantics requires a sequence of probability assignments rather than a single probability assignment. The semantics of this sequence is not clear.

Another concrete interpretation for belief is supplied by Levesque's work on making "believers" out of computers [38,40]. Levesque provides a functional view of knowledge bases, treating them as abstract data types on which two operations are performed: TELL and ASK. TELL adds new information to the database, and ASK is used to query the database. Levesque examines two languages in which these operations can be performed. One is a first-order language, and the other contains knowledge operators as well. Levesque shows that the stronger TELL and ASK operations, relying on the more expressive language, can be implemented using a first-order language only.

There are interesting similarities between Levesque's functional view of knowledge bases and our view of agents. Levesque's abstract definition of knowledge bases ignores the underlying implementation of these structures and defines them in term of their behavior—the responses they make to queries. This is quite similar to our functional view of agents emphasizing their behavior: Levesque's knowledge bases can be viewed as restricted agents having three actions—YES, NO, UNKNOWN—corresponding to their possible responses to queries. However, the goals of most situated systems (computer, mechanical, or biological) involve more than correctly representing their environment, although that is definitely useful, and their repertoire of actions is different than query responses. Our work examines this more general class of agents. Levesque's representation result, though somewhat different, provides conditions under which it is possible

to model an abstract knowledge base using first-order logic. Our existence results show when it is possible to model an agent using beliefs, preferences and a decision criterion.

Levesque's functional approach enables knowledge representation researchers to treat different knowledge-representation structures in a uniform manner, abstracting away implementation details that are irrelevant to their query answering behavior. This abstract view has proven extremely fruitful for understanding central issues in knowledge representation [39]. Given the success of Levesque's approach in the more limited domain, we hope that the more general abstraction given here will serve the same role in the general study of agents.

Plan recognition

In plan recognition [3, 14, 27, 30, 48] one tries to infer the plans of other agents by communicating with them or observing their behavior. This modeling task closely resembles the prediction task we discussed in Section 4, and the latter can be viewed as giving a semantic account of plan recognition. Most often, the modeled agent is human and often its acts are speech acts. Hence, this field does not attempt to provide a general semantic theory of mental-level models as abstract description of agents and devices. Grounding is not a central issue, since human agents (presumably) have an explicit mental state. The issue of existence is not dealt with either, but Kautz's use of circumscription [30] can be viewed as addressing the issue of model choice.

Given these general differences, among the work on plan recognition we find Pollack's work [48] to be the most relevant to our work. Pollack explicitly treats plans as complex mental attitudes involving beliefs and intentions of agents with goals. One of her points is that the planning agent's beliefs should be taken into account in deducing its plans, since they together with its goals determine its actions. This is reminiscent of our use of plausible outcomes. In addition, Pollack's model incorporates intentions which are missing from our model. However, Pollack's treatment is more syntactic than ours, and she presents various syntactic rules for plan inference. In addition Pollack does not deal with the issue of choice under uncertainty. She implicitly considers only 0/1 value functions and her agents always choose actions that achieve the goal given their beliefs. (This is akin to using the maximin criteria in our case.) Also, Pollack does not deal with the issue of belief change explicitly.

Machine learning

One driving force behind our research is the desire to use agent models for prediction in multi-agent systems. In machine learning, models are also constructed to help make predictions. For example, decision trees [50] provide a way of modeling observed instances in a manner that enables predicting the classes of future instances. Our work brings a special *bias* to the task of predicting agents' behavior, 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. It is an experimental question whether this bias is justified, but we believe that it is quite sensible. One of the reasons we consider the question of model existence to be important is because by answering it we can understand this bias better.

9. Discussion

In this paper we formalized a method for representing the state of an agent using mental attributes. McCarthy advocated this idea in [43] where he motivated this approach and suggested a number of ideas for formalizing it. Newell also advocated this idea, stressing the need for an abstract level of representation of programs and machines, which he called the *knowledge level* [45]. However, the tone of both papers is, in general, intuitive, informal, and motivational. In fact, in [46], Newell laments the lack of attempts to pursue this approach within the "logicist community". We believe this to be the first work within AI to provide formal treatment of these ideas. Our approach makes a number of semantic contributions: We proposed a structure for a mental-level description of an agent's state which makes explicit the agent's approach to choice under uncertainty, via its decision criterion. We explained how this high-level description is grounded in less abstract aspects of the agent, its actions. We advocated a holistic approach to the issue of grounding, in which one aspect of the agent's mental state, such as its beliefs, receives its meaning in the context of other mental attributes, such as preferences, and the whole state is related to the agent's behavior. We then investigated the properties of these models, showing a class of agents that can be modeled using this approach, and suggested two criteria for choosing among different consistent models.

In this work, we were not explicitly concerned with implementation issues. Rather, our goal was to clarify basic semantic questions. Despite considerable progress that previous research has made toward an understanding of formal semantic models of mental states, we felt that clearer and more complete treatment of the semantic foundations of mental-level models was needed. However, we now wish to briefly discuss representational issues pertaining to mental-level modeling. It should be noted that there have been important attempts to automate the design of agent models from which we can learn about these and other implementation issues. In particular, work on plan recognition, discussed earlier, is of considerable interest.

A primary concern in applications is what knowledge representation structures should be used to construct and store mental-level models. While it is natural and common to study the semantics of mental states in terms of possible worlds, this would not be a suitable way of actually representing these mental states.[23] The choice of the set of possible worlds is one of the framing decisions that the modeler must make. Usually, these worlds will be truth assignments to some set of propositions deemed relevant by the modeler. This points to the various logics for representing mental states as natural tools for reasoning about these models. In particular, Boutilier's logic for qualitative decision theory [6] seems promising. It is able to deal with beliefs, preferences, and a limited set of decision criteria; and its semantics is close to ours. Similarly, algorithms for constructing mental-level models will be needed. We have designed such algorithms, which appear in [9], but they are based on the state space representation and are only used for conceptual understanding of this issue.

[23] Semantically, probability distributions are also defined over sets of possible worlds, yet efficient representations of probability measures employ, e.g., Bayesian nets [47]. Similar structures developed for discrete notions as well, e.g.,[15], may be of use here.

In fact, a declarative implementation of the abstract models described in this paper may not be desirable. Indeed, an important point behind the concept of mental-level models, one stressed by Newell [45] in his discussion of the knowledge level, is that they provide an *abstract* description of a system. If we push this line of reasoning another step, we realize that the modeling process itself need not employ an abstract logical language, but can simply use the ideas discussed here and in similar work as abstract specifications. The implemented system will perform mental-level modeling of other agents using data-structures and algorithms that are suitable for its domain. A case in point are some of the recent systems used in the air-combat modeling domain [53, 59]. The goal of workers in these areas is to provide realistic air-combat simulators. Combat pilots use the current behavior of their opponents to ascribe them with goals and intentions. Then, they use these models to predict the future behavior of their opponents. Consequently, one would expect a simulated pilot to carry out such modeling activity. These problems are complicated by the fact that air-combat usually involves group activity and coordination.

The underlying semantic models used by these authors differ from ours in certain respects, placing greater emphasis on intention ascription which seems essential for generating realistic predictions. However, mental-level models play the above role of abstract specification tools. And while the implemented system is guided by the semantic model, the implementation itself makes extensive use of domain specific information and domain specific heuristics. For example, Tambe's system [59] assigns heuristic values to each consistent model as a means of choosing between different models consistent with the modeler's current information.

Indeed, often there are many ways to model a given behavior. Therefore, general background information about the agent modeled or agents of its type is important for discriminating among different possible models. For example, belief ascription is a modeling problem constrained by background knowledge of the agent's goals and decision criterion.[24] It will be the role of the modeler or its designer to supply this information. Alternatively, the modeler may gradually learn this information.

Another task which we considered as part of the specification of the model frame was that of identifying the set of local states of the agent. What's problematic about this task is that it seems to require access to the internal state of the agent. However, we believe that the following approach can be used to address this problem to a certain extent: identify local state changes with an action taken by the agent or new observation made by the agent. Of course, in principle, recognizing observations that are made by the agent also requires access to its internal state, but reasonable deductions can be made in many cases. Otherwise, we simply use the agent's actions as indications of local state change. As one would expect, the less we know about the agent and its sensory capabilities, the harder this problem is and the harder it is to construct a model frame.

[24] One reason for stressing belief ascription, which requires this type of knowledge, is that agents can be equipped with such knowledge. We believe that an agent's goals and its approach to decision making are somewhat stable aspects of it, determined by its designer. Therefore, through observations over time, a reasonable understanding of them can be attained. However, dynamic agents that learn and make observations will have much less stable beliefs.

Various issues remain for future work. Indeed, much more research on efficient representation, construction, and reasoning techniques is required. Refinements of the model structure are also needed. In particular, a realistic model must somehow deal with the bounded computational resources available to real agents. The notion of intention may provide some ability to deal with bounded rationality, as possibly could new decision criteria. However, considering the fact that, unlike humans, artificial agents are likely to have diverse architectures, it is unlikely that a single choice of decision criterion will be adequate for modeling all agents. This, again, points to the importance of more existence results that will give us a better sense of the modeling capabilities of different proposed models. Finally, as observed by researchers on plan recognition, techniques for choosing among candidate models are of great practical importance and they should be studied farther. We hope to pursue some of these questions in our future work.

Acknowledgment

We are grateful to Yoav Shoham for important discussions regarding this work, to Joe Halpern for extensive comments and suggestions on previous drafts of this paper, and to Craig Boutilier, the anonymous referees, Alvaro del Val, Bruce Donald, Nir Friedman, and Daphne Koller for their useful comments and suggestions on earlier versions of this work. We also wish to thank Robert Aumann, David Kreps, and Dov Monderer for helping us put this work in perspective.

Parts of this work were conducted while the authors were at Stanford University, partially supported by grants from the National Science Foundations, Advanced Research Projects Agency and the Air Force Office of Scientific Research.

Appendix A. Proofs

Theorem 21. *Assuming perfect recall, a belief assignment B is admissible iff there is a ranking function r (i.e., a total pre-order) on the possible initial worlds such that* $B_I(l) = \{s \in PW_I(l) \mid s \text{ is } r\text{-minimal in } PW_I(l)\}$.

Proof. For one direction, assume we are given a ranking, we must show that it induces admissible beliefs. First, notice that because of perfect recall, the set of possible initial worlds can only decrease. This means that any world that was minimal among the possible worlds at one point, will always be minimal in the future, as long as it is still possible. Let l and l' be consecutive local states on some run. The previous observation implies $B_I(l') \supseteq B_I(l) \cap PW_I(l')$. However, when $B_I(l) \cap PW_I(l') \neq \emptyset$ then no world that was not minimal in $PW_I(l)$ can become minimal in $PW_I(l')$. Hence, $B_I(l') = B_I(l) \cap PW_I(l')$.

For the other direction, let I be the set of initial possible worlds and assume that B_I is admissible, we construct a ranking function based on B_I. First, notice that for any two initial local state l_1, l_2, the sets $PW_I(l_1)$ and $PW_I(l_2)$ are disjoint (because of perfect recall). Furthermore, perfect recall implies that, if l_1' is a local state that can be reached

from a state in $PW_I(l_1)$ and l_2' is a local state that can be reached from a state in $PW_I(l_2)$, then $PW_I(l_1')$ and $PW_I(l_2')$ are disjoint. Therefore, we can separately rank all states reachable from $PW_I(l_1)$ and $PW_I(l_2)$ and then unite these ranking in any manner that preserves the ranking over each set. Therefore, without loss of generality, we will assume that there is only one possible initial state l and show how to rank $PW_I(l)$. At the nth step of the algorithm we assign (a possibly empty) set to the nth rank. At l we assign $B_I(l)$ to rank 1, the lowest (most normal) rank. Let $l_1, \ldots, l_k$ be the possible local states at time n in all runs commencing in $PW_I(l)$. We assign to rank n all those states in $B_I(l_j)$ $(1 \leqslant j \leqslant k)$ that have not been assigned a rank so far.

The ranking we defined defines an admissible belief assignment, and we now have to show that it is identical to the original belief assignment. We prove this by induction on the time at which a local state appears in a run. It clearly does at the initial local state l. Let l' be some local state that appears at time n on some run and let l'' be one of its children (i.e., l'' is a local state at time $n+1$ at some run whose local state at time n was l'). By the induction hypothesis, $B_I(l')$ is indeed equal to the set of minimal ranked worlds in $PW_I(l')$. Suppose that $B_I(l'')$ contains a state that is not in $B_I(l')$. Because B_I is admissible, $B_I(l') \cap PW_I(l'') = \emptyset$. Consequently, $B_I(l') \cap B_I(l'') = \emptyset$. Moreover, for any $\hat{l}$ that may have preceded l'', it is the case that $B_I(\hat{l}) \cap PW_I(l'') = \emptyset$. Otherwise, using the induction hypothesis, some state in $PW_I(l'')$ would have been minimal before, and hence would have been in $B(l')$. This means that none of the states in $PW_I(l'')$ would have been assigned a rank of n or lower. In stage $n+1$ of the construction process we assigned all of $B_I(l'')$ the rank of $n+1$ while the rest of $PW_I(l'')$ has not yet been assigned a rank. Therefore, the minimally ranked elements of $PW_I(l'')$ are precisely $B_I(l'')$. If $B_I(l'') \cap B(l') \neq \emptyset$ then we know that $B_I(l'') = PW_I(l'') \cap B(l')$. Since perfect recall implies $PW_I(l'') \subseteq PW_I(l')$, we get that $B_I(l'')$ are the minimal worlds in $PW_I(l'')$ according to the ranking. □

Theorem 24. *Let $\mathcal{P}$ be the protocol of an agent. If this agent 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.*

Proof. The beliefs ascribed at the initial states uniquely determine the agent's beliefs at subsequent non-revision states assuming admissibility. We claim that this belief assignment is consistent with observed behavior. Suppose not, this means that at some local state another protocol $\mathcal{P}'$ would be chosen given these beliefs. (That is, its plausible outcome would be strictly more preferred than that of the observed protocol). However, using the assumptions of perfect recall, the sure-thing principle, the fact that ρ is static with respect to to u, and that the beliefs are admissible, we will show that if a protocol $\mathcal{P}'$ is more preferred at this state, it would have been more preferred at the previous state as well. By repeating this process, we get that this protocol would have been more preferred initially, contradicting the fact that the beliefs assigned at the initial state are consistent with the observed protocol.

In order to see this, let l' and l'' be the children of a local state l (the same argument applies when there are more children), that is the possible local states that can follow l when we perform at l the action assigned by the actual protocol, which by our

assumption, is most preferred at l. By definition, we have that $PW_I(l') \cap PW_I(l'') = \emptyset$. If l'' is a revision state then we can conclude that the agent's beliefs in l and l' (i.e., its B_I) are identical (based on admissibility). This together with the fact that ρ is static with respect to u implies that the same protocols would be preferred in both local states. By the same protocols we mean, the same from l' on, since there is no point in comparing protocols on l' that take a different action on l than the actual protocol, since they would not lead to l'. Next, suppose that l'' is not a revision state. Consider the protocol P' that is identical with the actual protocol up to and including l, and which assigns at l', l'', and their descendents the protocols most preferred at l' and l'', respectively. These two protocols really apply to different local states, since no local state is reachable from both l' and l'' (because of perfect recall). Therefore, there is not conflict. We know that P' is most preferred in l' given $B_I(l')$. Hence, using the fact that ρ is static with respect to u, we get that P' would be most preferred in l among those protocols that are identical to P' on l, if the beliefs in l were exactly $B_I(l')$. A similar argument applies to l''. Now, we know that $B_I(l) = B_I(l') \cup B_I(l'')$ and that $B_I(l') \cap B_I(l'') = \emptyset$. Therefore, we can apply the sure-thing principle to show that P' would be most preferred according to $B_I(l)$ among the set of protocols that are identical to P' up to and including l. In particular, this class of protocols contains the protocols most preferred at l (by our assumption). Hence P' is most preferred at l. □

Theorem 27. *Let $\mathcal{A}$ be an agent whose possible runs have bounded length, for which we can ascribe a consistent mental-level model $\langle \mathcal{L}_{\mathcal{A}}, A_{\mathcal{A}}, B_I, u, \rho \rangle$, where B_I is admissible and ρ satisfies the sure-thing principle. Then, $\mathcal{A}$ can be ascribed a consistent decomposed mental-level model $\langle \mathcal{L}_{\mathcal{A}}, A_{\mathcal{A}}, B_{cur}, u_{cur}, \rho \rangle$.*

Proof. Theorem 3 of [11] says that when an agent can be ascribed admissible beliefs and the assumptions we have made in this paper are satisfied, i.e., ρ satisfies the sure-thing principle and is static with respect to u, and its runs have bounded length, then its protocol can be derived from its mental-level model by using backwards induction (BI). (We repeat the proof next.) Hence we can treat its protocol as a backwards induction protocol. This means that at each point, the agent can be viewed as choosing the best action (according to its beliefs), given its choices for the following actions. We first transform B_I into B_{cur}, which is a cosmetic change. Next, we assign to each state as its value, the value of the run suffix obtained by running the backwards induction protocol from this point on. Since backwards induction chooses the best action given its future choices, which are now embodied in the value of each state, we get equivalent choices.

Proof of the relationship with backwards induction: Suppose that an agent can be ascribed admissible beliefs. This means that it has a mental-level model with admissible beliefs according to which its actual protocol is most preferred at all states, including the initial state. We now show that $\mathcal{P}$ is a backwards induction protocol for $\mathcal{A}$ iff it is most preferred at the initial state, given our assumptions on ρ and u and the assumption of admissible beliefs. One direction of this claim will then give us what we need for the above theorem. In order to simplify the statement of the proof, we will use here a stronger notion of most preferred protocol, where whenever $\mathcal{P}$ and $\mathcal{P}'$ are equally preferred in local state l, but $\mathcal{P}$ is more preferred in some revision state that can be

reached by both protocols, we will consider $\mathcal{P}$ to be more preferred at l as well. This will apply when we have two protocols that behave identically on all non-revision states, but diverge on revision states. The plausible outcome of such protocols is identical, and the decision criterion will not distinguish between them. However, once we get to the revision state, one will be preferred over the other, and so we use that fact already at the initial state.

First we show that the protocol most preferred at the initial local state is a BI protocol. We do so by induction. For tree of depth 1 this is obvious. Suppose that we have shown this to be the case for trees of depth less than n and let *root*, the initial local state, be of height n. Suppose that $\mathcal{P}$ is most preferred on *root*, to which it assigns the actions a. By the previous theorem $\mathcal{P}$ is also most preferred at its children. Without loss of generality, assume that it has two children l' and l''. By the induction hypothesis, the protocols most preferred on l' and l'' are BI protocols. Suppose that the BI protocol assigns b to *root*. By its definition, the BI protocol selects the action most preferred given the BI choices for the children. If it chose b at *root*, then this means that doing b and then the BI protocol on B's children is better than doing a and then the BI protocol on a's children. But this contradicts the fact that $\mathcal{P}$ is most preferred at *root*. Therefore, $\mathcal{P}$ is a BI protocol.

We must now show that BI protocols are most preferred at the initial state. Suppose that $\mathcal{P}$ is BI at *root*. This means that it is a most preferred protocol among BI protocols at *root*. Since we have shown that the most preferred protocols at *root* are BI, then $\mathcal{P}$ is at least as preferred as them, and hence, is most preferred. □

Corollary 28. *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 value function* over states *such that its observed* action *has a most preferred plausible outcome according to the local value function.*

Proof. This is an immediate conclusion from the previous two theorems. The first allows us to ascribe a structure with admissible beliefs, while the latter allows us to switch to a local representation in that case. □

Theorem 34. *Given a belief ascription problem with a decision criterion that is closed under unions, if a consistent (explanatory) belief assignment exists then there is a unique minimal (explanatory) consistent belief assignment.*

Proof. The belief assignment has to satisfy local criteria, i.e., for any local state the actual protocol has a most preferred plausible outcome with respect to the decision criteria. Since the decision criterion is closed under unions, then we have that if B and B' are consistent belief assignments then so is $B \cup B'$. This ensures uniqueness. □

Theorem 37. *For agents with perfect recall, if the decision criteria is closed under unions then the set of consistent admissible (explanatory) belief assignments, if non-empty, contains a unique minimal (explanatory) belief assignment.*

Proof. Given that the agents have perfect recall we know that we can think of their belief assignments as rankings. Assume that B_1 and B_2 are two different belief assignments, none of which is more general than the other. (Throughout this proof, the belief assignments we refer to assign subsets of $PW_I(\cdot)$.) We show that we can construct an admissible belief assignment that is more general than both. Let l be a local state. Without loss of generality assume that $B_1^0 \neq B_2^0$. (We use B^i to denote the states in the ith rank of B). We claim that an admissible belief assignment B exists such that $B^0 = B_1^0 \cup B_2^0$.

We let B be the union of B_1 and B_2 on the initial local state. Thus, $B^0 = B_1^0 \cup B_2^0$. Because the decision criterion is closed under unions, we know that that's fine for these states. The definition of admissibility then implies the value of B on any state l for which $PW_I(l)$ contains a state in B^0. This is easily seen to be consistent, since it implies that $B(l)$ is either $B_1(l), B_2(l)$ or their union, all of which are consistent with the protocol. Next, we look to revision states according to B. In these states none of B_1^0 or B_2^0 are possible, hence we can again assign the union of B_1 and B_2 on these states. This process continues until all states are assigned, and as was seen, at each point B is consistent with the protocol. □

Theorem 38. *If A is a goal-seeking agent then it can be ascribed a unique minimal admissible belief assignment and a unique minimal explanatory admissible belief assignment.*

Proof. We construct an admissible belief assignment B_I. Let l^i be the initial local state. If $\mathcal{P}$ performs HALT at l^i then either it is a goal state and we choose $B_I(l^i) = PW_I(l^i)$ or otherwise there is a world $s \in PW_I(l^i)$ such that the goal cannot be reached from s (using Rational Despair). We let $B_I(l^i)$ be the set of all such worlds. If $\mathcal{P}$ does not perform HALT we choose the maximal set, S, of states under which $\mathcal{P}$ is not weakly dominated by any other protocol. A maximal set exists because the decision criterion is closed under unions. Such a set is not empty because otherwise $\mathcal{P}$ must be HALT (applying Rational Effort). We let $B_I(l^i) = S$ By definition of admissibility, for any state l consistent with S (i.e., $S \cap PW_I(l) \neq \emptyset$) we must define $B_I(l) = S \cap PW_I(l)$, therefore, we need to see that in any state l consistent with S, $\mathcal{P}$ is still not weakly dominated according to $S \cap PW_I(l)$. Assume the contrary. This means that for some state $s \in S \cap PW_I(l)$, $\mathcal{P}$ does not achieve the goal, while some other protocol, $\mathcal{P}'$, does achieve the goal. But this means that there is some protocol $\mathcal{P}''$ such that $\mathcal{P}''$ is the same as $\mathcal{P}$ up to l, and the same as $\mathcal{P}'$ from l. $\mathcal{P}''$ weakly dominates $\mathcal{P}$ in l^i. This contradicts our choice of $\mathcal{P}$ in l^i.

In states l not consistent with S (i.e., states in which $S \cap l = \emptyset$), we cannot achieve the goal using $\mathcal{P}$. By the Rational Effort postulate this means that at these local states the protocol must be HALT (since a protocol that cannot achieve the goal in any state does not weakly dominate HALT). By the Rational Despair postulate this means that there is some world $s \in PW_I(l)$ from which no protocol attains the goal. We let $B_I(l)$ be the set of all such worlds.

To obtain the result, but using explanatory belief assignment, we must construct an explanatory belief assignment. Uniqueness follows from the fact that the decision

criterion used is closed under unions. The construction is as above, but instead we define S to be the set of states on which the protocol achieves the goal.

Finally, notice that if we drop the uniqueness requirement then we can allow any decision criterion consistent with weak dominance (i.e., one in which if w weakly dominates v than w is strictly more preferred than v). □

Corollary 39. *An agent that can be modeled as a B-type agent can be viewed as a plausible outcome maximizer that uses some admissible belief assignment and a decision criterion consistent with weak dominance.*

Proof. It is easy to check that B-type agents satisfy both of our rationality postulate. Thus, they differ from goal-seeking agents only in their decision criterion. Using the remark in the previous proofs, we see that they can be ascribed admissible beliefs. □

References

[1] E. Adams, *The Logic of Conditionals* (Reidel, Dordrecht, Netherlands, 1975).

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

[3] J.F. Allen, Recognizing intentions from natural language utterances, in: M. Brady and R.C. Berwick, eds., *Computational Models of Discourse* (MIT Press, Cambridge, MA, 1983).

[4] F. Bacchus, A.J. Grove, J.Y. Halpern and D. Koller, Statistical foundations for default reasoning, in: *Proceedings IJCAI-95*, Montreal, Que. (1995) 563–569.

[5] D.G. Bobrow, Artificial intelligence in perspective: a retrospective on fifty volumes of the *Artificial Intelligence Journal*, *Artificial Intelligence* 59 (1993) 5–20.

[6] C. Boutilier, Toward a logic for qualitative decision theory, in: J. Doyle, E. Sandewall and P. Torasso, eds., *Proceedings 4th International Conference on Principles of Knowledge Representation and Reasoning*, Bonn (1994) 75–86.

[7] C. Boutilier and M. Goldszmidt, Revising by conditional beliefs, in: *Proceedings AAAI-93*, Washington, DC (1993) 648–654.

[8] R.I. Brafman, Qualitative models of information and decision making: Foundations and applications, Ph.D. Thesis, Stanford University, Stanford, CA (1996).

[9] R.I. Brafman and M. Tennenholtz, Belief ascription, Working notes (1994).

[10] R.I. Brafman and M. Tennenholtz, Belief ascription and mental-level modeling, in: J. Doyle, E. Sandewall and P. Torasso, eds., *Proceedings 4th International Conference on Principles of Knowledge Representation and Reasoning*, Bonn (1994) 87–98.

[11] R.I. Brafman and M. Tennenholtz, Towards action prediction using a mental-level model, in: *Proceedings IJCAI-95*, Montreal, Que. (1995) 2010–2016.

[12] R.I. Brafman and M. Tennenholtz, On the foundations of qualitative decision theory, in: *Proceedings AAAI-96*, Portland, OR (1996) 1291–1296.

[13] R.I. Brafman and M. Tennenholtz, On the axiomatization of qualitative decision criteria, in: *Proceedings AAAI Spring Symposium on Qualitative Preferences in Deliberation and Practical Reasoning* (1997) 29–34.

[14] E. Charniak and R.P. Goldman, A Bayesian model of plan recognition, *Artificial Intelligence* 64 (1993) 53–80.

[15] A. Darwiche and J. Pearl, Symbolic causal networks, in: *Proceedings AAAI-94*, Scattle, WA (1994) 238–244.

[16] A. del Val and Y. Shoham, Deriving properties of belief update from theories of action, in: *Proceedings IJCAI-89*, Detroit, MI (1989) 584–589.

[17] T.G. Dietterich, Machine learning, in: *Annual Review of Computer Science*, Vol. 4 (Annual Reviews Inc., Palo Alto, CA, 1990) 255–306.
[18] J. Doyle, Reasoned assumptions and Pareto optimality, in: *Proceedings IJCAI-85*, Los Angeles, CA (1985) 87–90.
[19] J. Doyle, Constructive belief and rational representation, *Comput. Intell.* 5 (1989) 1–11.
[20] R. Fagin, J.Y. Halpern, Y. Moses and M.Y. Vardi, *Reasoning about Knowledge* (MIT Press, Cambridge, MA, 1995).
[21] P.C. Fishburn, *Nonlinear Preference and Utility Theory* (Johns Hopkins University Press, Baltimore, MD, 1988).
[22] N. Friedman and J.Y. Halpern. A knowledge-based framework for belief change, Part I: Foundations, in: *Proceedings 5th Conference on Theoretical Aspects of Reasoning about Knowledge*, San Francisco, CA (Morgan Kaufmann, Los Altos, CA, 1994).
[23] P. Gärdenfors and N.E. Sahlin, eds., *Decision, Probability and Utility* (Cambridge University Press, New York, 1988).
[24] 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: Proceedings 3rd International Conference (KR-92)*, Cambridge, MA (1992) 661–672.
[25] A.J. Grove, Two modellings for theory change, *J. Philosophical Logic* 17 (1988) 157–170.
[26] J.Y. Halpern and Y. Moses, Knowledge and common knowledge in a distributed environment, *J. ACM* 37 (1990) 549–587.
[27] M.J. Huber, E.H. Durfee and M.P. Wellman, The automated mapping of plans for plan recognition, in: R.L. de Mantaras and D. Poole, eds., *Uncertainty in AI, Proceedings 10th Conference* (1994) 344–352.
[28] H. Katsuno and A. Mendelzon, On the difference between updating a knowledge base and revising it, in: *Principles of Knowledge Representation and Reasoning: Proceedings 2nd International Conference (KR-91)*, Cambridge, MA (1991) 387–394.
[29] H. Katsuno and A.O. Mendelzon, Propositional knowledge base revision and minimal change, *Artificial Intelligence* 52 (1991) 263–294.
[30] H.A. Kautz, Generalized plan recognition, in: *Proceedings AAAI-86*, Philadelphia, PA (1986).
[31] S. Kraus and D.J. Lehmann, Knowledge, belief and time, *Theoret. Comput. Sci.* 58 (1988) 155–174.
[32] D.M. Kreps, *Notes on the Theory of Choice* (Westview Press, Boulder, CO, 1988).
[33] D.M. Kreps, *A course in Microeconomic Theory* (Princeton University Press, Princeton, NJ, 1990).
[34] D.M. Kreps and R. Wilson, Sequential equilibria, *Econometrica* 50 (1982) 863–894.
[35] S. Kripke, Semantical considerations of modal logic, *Z. Math. Logik Grundlag. Math.* 9 (1963) 67–96.
[36] P. Lamarre and Y. Shoham, Knowledge, certainty, belief and conditionalization, in: *Proceedings 4th International Conference on Principles of Knowledge Representation and Reasoning*, Bonn (1994).
[37] D. Lehmann and M. Magidor, What does a conditional knowledge base entail?, *Artificial Intelligence* 55 (1992) 1–60.
[38] H.J. Levesque, Foundations of a functional approach to knowledge representation, *Artificial Intelligence* 23 (1984) 155–212.
[39] H.J. Levesque, Knowledge representation and reasoning, in: *Annual Review of Computer Science*, Vol. 1 (Annual Reviews Inc., Palo Alto, CA, 1986) 255–287.
[40] H.J. Levesque, Making believers out of computers, *Artificial Intelligence* 30 (1986) 81–108.
[41] R.D. Luce and H. Raiffa, *Games and Decisions* (John Wiley & Sons, New York, 1957).
[42] M. Machina, Dynamic consistency and non-expected utility models of choice under uncertainty, *J. Economic Literature* 27 (1989) 1622–1668.
[43] J. McCarthy, Ascribing mental qualities to machines, in: M. Ringle, ed., *Philosophical Perspectives in Artificial Intelligence* (Humanities Press, Atlantic Highlands, NJ, 1979).
[44] S. Morris, Revising knowledge: a hierarchical approach, in: *Proceedings 5th Conference on Theoretical Aspects of Reasoning about Knowledge*, San Francisco, CA (Morgan Kaufmann, Los Altos, CA, 1994).
[45] A. Newell, The knowledge level, *AI Mag.* 2 (2) (1981) 1–20.
[46] A. Newell, Reflections on the knowledge level, *Artificial Intelligence* 59 (1993) 31–38.
[47] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, Palo Alto, CA, 1988).

[48] M.E. Pollack, Plans as complex mental attitudes, in: P.R. Cohen, J. Morgan and M.E. Pollack, eds., *Intentions in Communication* (MIT Press, Cambridge, MA, 1990) 77–104.
[49] M.E. Pollack, D.J. Israel and M. Bratman, Towards an architecture for resource-bounded agents, Technical Report, Technical Note 425, SRI International, Menlo Park, CA (1987).
[50] J.R. Quinlan, Induction of decision trees, *Machine Learning* 1 (1986) 81–106.
[51] A.S. Rao and M.P. Georgeff, Deliberation and its role in the formation of intentions, in: *Proceedings 7th Annual Conference on Uncertainty Artificial Intelligence (UAI-91)*, Los Angeles, CA (1991).
[52] A.S. Rao and M.P. Georgeff, An abstract architecture for rational agents, in: *Principles of Knowledge Representation and Reasoning: Proceedings 3rd International Conference (KR-92)*, Cambridge, MA (1992) 439–449.
[53] A.S. Rao and G. Murray, Multi-agent mental state recognition and its application to air-combat modeling, in: *Proceedings 13th International DAI Workshop (DAI-13)*, Seattle, WA (1994).
[54] S.J. Rosenschein, Formal theories of knowledge in AI and robotics, *New Generation Comput.* 3 (1985) 345–357.
[55] S.J. Rosenschein and L.P. Kaelbling, A situated view of representation and control, *Artificial Intelligence* 73 (1995) 149–174.
[56] L.J. Savage, *The Foundations of Statistics* (Dover, New York, 1972).
[57] Y. Shoham, Agent-oriented programming, *Artificial Intelligence* 60 (1993) 51–92.
[58] Y. Shoham and S.B. Cousins, Logics of mental attitudes in AI: a very preliminary survey, in: G. Lakemeyer and B. Nebel, eds., *Foundations of Knowledge Representation and Reasoning* (Springer, Berlin, 1994).
[59] M. Tambe, Recursive agent and agent-group tracking in real-time, dynamic environments, in: *Proceedings International Conference on Multi-Agent Systems (ICMAS-95)*, San Francisco, CA (1995).
[60] R.H. Thomason, Towards a logical theory of practical reasoning, in: *AAAI Spring Symposium on Reasoning About Mental States: Formal Theories and Applications* (1993).
[61] G. Tidhar, Personal communication (October 1996).
[62] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, NJ, 1944).