Ascribing Beliefs

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

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

Abstract

Models of agents that employ formal notions of mental states are useful and often easier to construct than models at the symbol (e.g., programming language) or physical (e.g., mechanical) level. In order to enjoy these benefits, we must supply a coherent picture of mental-level models, that is, a description of the various components of the mental level, their dynamics and their inter-relations. However, these abstractions provide weak modelling tools unless (1) they are grounded in more concrete notions; and (2) we can show when it is appropriate to use them. In this paper we propose a model that grounds the mental state of the agent in its actions. We then characterize a class of *goal-seeking* agents that can be modelled as having beliefs.

This paper emphasizes the task of belief ascription. On one level this is the practical task of deducing an agent's beliefs, and we look at assumptions that can help constrain the set of beliefs an agent can be ascribed, showing cases in which, under these assumptions, this set is unique. We also investigate the computational complexity of this task, characterizing a class of agents to whom belief ascription is tractable. But on a deeper level, our model of belief ascription supplies concrete semantics to beliefs, one that is grounded in an observable notion – action.

1 INTRODUCTION

Abstractions play an important role in our reasoning ability. Arguably, the most fundamental abstraction we use involves modelling other entities as having mental states. We use it to model other biological entities and perhaps even ourselves; it may even be used in modelling complex mechanical entities. Indeed, Allen Newell, in a famous paper [Newell, 1980], argues that intelligent systems can be (approximately) described at a level higher than the symbol (e.g., programming language) level and the physical level, which he calls the *knowledge level*.

Having a mental-level model offers many advantages. First, it allows us to describe a system's behavior without a detailed description of its lower-level, e.g., its implementation as machine code, or its physical components. A mental-level model is also much more accessible and intuitive to us. We can, therefore, use it to critique a system by looking at its beliefs and asking ourselves whether they make sense. Similarly, we can examine a system's goals and criticize them. And while a model at the symbol or physical level requires detailed knowledge that is often not available, we can usually construct a mental model of an agent by observing its behavior or by using general knowledge about the typical behaviors of this agent. An understanding of the way this behavior is implemented within this agent is not necessary, as we know from our experience. This makes possible the task of predicting an agent's behavior without access to its program. And as John McCarthy says [McCarthy, 1979],

> (Ascription of mental states) is useful when the ascription helps us understand the structure of the machine, its past or future behavior, or how to repair and improve it. It is perhaps never required even for humans, but expressing reasonably briefly what is actually known about the state of a machine ... may require ascribing mental qualities.

In order to use these abstractions we must provide the foundations required for modelling agents at the mental level. First, we need good models of the mental level, i.e., its components, the way they interact and the way they change over time. Yet, without grounding these abstract notions in more concrete ones, we are walking on thin ice, because it is not clear what notions such as belief mean, nor when is it appropriate to use them. And even those who may argue for the 'objective' existence of epistemic states in some entities, are on shaky ground when it comes to modelling computer systems and mechanical devices. The abstract semantics for belief using Kripke structures [Kripke, 1963, Hintikka, 1962], can only serve as the first step towards an understanding of such notions.

In this paper we propose a model that grounds the mental state of an agent in its actions. We develop a formal model of the mental level, which is motivated by work in decision theory [Luce and Raiffa, 1957] and the work of [Rosenschein, 1985] and [Halpern and Moses, 1990] on knowledge ascription. The model is quite simple and intuitive. It uses a number of components: beliefs, utilities, and a decision-strategy to construct a mental-level model. This model relates these components among themselves and with the agent's behavior, i.e., its choice of action. It is built upon a lower-level description of the agent, which we will call the *physical level*. The relation of the various components at the mental level and the agent's behavior is embodied by the *agency hypothesis*, which implicitly defines the states an agent believes in as those states that affect its choice of action. This manner of grounding the notion of beliefs in the agent's (observable) choice of action provides a way of addressing the question of *adequacy*, i.e., when can we actually use mental-level modelling. The answer can be given through a characterization of classes of behaviors that make this type of modelling possible.

We start with a static model. In this model the agent associates with each possible action a number of plausible outcomes, which depend on the agent's beliefs. The agent assigns a utility to each outcome, representing the relative desirability of this outcome. The agent then uses its decision strategy to choose an action based on the utilities of this action's outcomes. We call the hypothesis that the the agent's (ascribed) mental states relates to its action in this manner, the *agency hypothesis*. The agency hypothesis grounds the agent's beliefs in its choice of action.

Based on the static model we develop a more dynamic model that also takes into account the issue of belief change, and provide two interesting representation theorems. These theorems relate certain patterns of belief change with a static representation of belief based upon partial and total pre-orders.

With this model at hand we proceed to specifically examine the problem of belief ascription. Our abstract view of belief ascription is that of constraining the (ascribed) beliefs of the agent by its (observable) choice of action via the agency hypothesis. In order to address the adequacy problem, we define a class of *goal-seeking* agents and show that they can be ascribed belief within our framework. Unfortunately, in practice, it is often the case that we cannot ascribe an agent unique beliefs based on the available information. We examine this issue and suggest two general heuristics for choosing among multiple candidates. We give conditions under which these criteria suggest unique beliefs. Another practical issue involves the computational complexity of belief ascription. We show that although in general it may be difficult (i.e., CO-NP-hard), for some agents it is tractable.

On one level, belief ascription is the task of deducing an agent's beliefs, and this paper attempts to address some practical issues that arise in performing this task. However, the deeper meaning of this work lies in the semantics of beliefs as modelling tools defined by the agent's choice of action.

1.1 A MOTIVATING EXAMPLE

To introduce the problem of belief ascription and the motivation behind our proposed solution, we present the following example.

Say we only care about four sets of worlds, described by the propositions *cold* and *rainy*. Our agent, Alice, has an accurate thermostat at home, but no windows. In a $cold \wedge rainy$ world, there are two worlds Alice considers possible: $cold \wedge rainy$ and $cold \wedge \neg rainy$. Because in all her possible worlds *cold* holds, Alice *knows* that it is cold. In general, to determine what Alice knows, we construct her set of possible worlds. [Halpern and Moses, 1990] shows us how we can construct this set given an appropriate description of Alice.

Alice does not *know* that it is rainy, but does she *believe* that it is rainy? It seems, that to answer this question, more information is required. So suppose we see Alice leaving home without an umbrella. This seems to indicate that she does not believe it is rainy, for otherwise she would have taken an umbrella. So based on Alice's action we have deduced her beliefs. However, to do so we implicitly assumed that Alice does not like getting wet and that she had the choice of taking an umbrella. That is, we used information regarding Alice's desires and possible choices of action.

Let's be more precise. The following matrix describes the outcome of Alice's two possible actions.

	rainy	$\neg$*rainy*
take umbrella	dry,heavy	dry,heavy,look stupid
leave umbrella	wet,light	dry,light

Suppose that Alice's preferences are described by the following utility function:

	rainy	$\neg$*rainy*
take umbrella	5	-1
leave umbrella	-4	10

A belief that $\neg rainy$ is the only plausible world would adequately explain Alice's behavior, as it will make the choice of leaving the umbrella the preferred one. Are other beliefs consistent with her behavior? Well, she could not believe *rainy* to be the only plausible world, for then she would have taken the umbrella. Could she consider both worlds plausible? The answer depends on her *decision criterion*. If she prefers to be on the safe side, employing a *maximin* strategy, which attempts to maximize the worst case outcome, then had she believed both worlds to be plausible, she would have taken the umbrella (with a worst case payoff of -1) rather than leaving it (with a worst case payoff of -4). But if Alice follows the *principle of indifference*, which takes the average payoff across plausible states, belief in both states is consistent, since leaving the umbrella has a better average payoff (3) than taking it (2).

Overview The next section describes a mental-level model based upon the notions of knowledge, belief, decision criteria, and utilities. In Section 3 this model is used to define belief ascription, and consequently, provide a definition of belief based on choice of action. As we will see, often we cannot ascribe unique beliefs to an agent, and in Section 4 we suggest how one can narrow the choice of appropriate belief ascriptions. In Section 5 we add time to the static model of Section 2, enabling us to investigate the issue of belief change in Section 6. In Section 7, having described a dynamic picture of the mental level, we characterize a class of agents to which belief can be ascribed using this model. In Section 8 we look at complexity issues, i.e., how difficult is belief ascription and we characterize a class of agents to whom belief ascription is tractable. In section 9 we return to the issue of choosing among belief assignments, characterizing a class of agents to whom the criteria of Section 4 enable narrowing the choice to a unique belief assignment. Section 10 concludes with a discussion of related work and some of our assumptions. Proofs of all theorems appear in the appendix.

2 THE FRAMEWORK

Starting with a physical level description of a system containing a single agent and an environment, we review knowledge ascription, following [Halpern and Moses, 1990]. Then, we introduce a number of new elements, beliefs, decision criteria, and utilities, and relate them to the agent's behavior. To make our definitions clear we will accompany them with a simplified version of McCarthy's famous thermostats example.

Example 1 *In [McCarthy, 1979], McCarthy shows how we often ascribe mental states to simple devices, thermostats in that case. Our goal is to formalize this informal discussion. 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.*

2.1 THE PHYSICAL LEVEL AND KNOWLEDGE

An *agent* is described by a set of possible (local) states and a set of possible actions. The agent functions within an *environment*, which may also be in one of a number of states. We refer to the state of the system, i.e., that of both the agent and the environment as a *global state*. W.l.o.g., we will assume that the environment does not perform actions. The effects of the agent's actions are a (deterministic) function of its state and the environment's state.[1] This effect is described by the *transition function*. Together, the agent and the environment constitute a state machine with two components, with transitions at each state corresponding to the agent's possible actions. It may be the case that not all combinations of an agent's local state and an environment's state are possible. Those global states that are possible are called *possible worlds*.

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

A context specifies the environment (since $L_{\mathcal{E}}$ is implicit in τ) and the effects of the agent's actions on the whole system. Later on, when we add time to the picture it will also specify the possible starting points of a system.

Example 1 *(continued): For our thermostat* $L_{\mathcal{A}} = \{-,+\}$. $-$ *corresponds to the case when the thermostat indicates a temperature that is less than the desired room temperature and* $+$ *corresponds to a temperature greater or equal to the desired room temperature. However, we take into account the fact that the thermostat may be mistaken in its measurement of the room's temperature, which is indeed one of the situations McCarthy considers. The thermostat's actions,* $A_{\mathcal{A}}$, *are* {turn-on, shut-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. Therefore the set of possible worlds is exactly* $L_{\mathcal{A}} \times L_{\mathcal{E}}$. *We chose the following transition function:*

[1] A framework in which the environment does act can be mapped into this framework using richer state descriptions and larger sets of states, a common practice in game theory.

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

	cold	*ok*	*hot*
turn-on	*ok*	*hot*	*hot*
shut-off	*cold*	*ok*	*ok*

In our example, the result of an action does not depend on the state of the thermostat. To simplify matters we assume that the thermostat is not affected by its actions, although this does not matter in this example.

Knowledge can be ascribed to the agent using the notion of a local state. An agent can distinguish between two worlds in S if and only if *its* state in them, is different. Therefore, an agent whose local state is l can rule out as impossible all worlds in which his local state would have been different, but cannot rule out worlds in S in which his local state would have been l. Knowledge corresponds to what holds in all worlds the agent cannot distinguish from the actual world.

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

Example 1 *(continued): While the thermostat, by definition, knows its local state, it knows nothing about the room's temperature. This stems from the fact that in our model we allowed for the possibility of a measurement error by the thermostat, making all elements of* $L_{\mathcal{A}} \times L_{\mathcal{E}}$ *possible, e.g.,* $(-, hot)$ *is a possible world.*

If truth assignments (for some given language) are attached to each world in S and 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.

The agent's observed, or programmed behavior is described by the protocol.

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

Example 1 *(continued): Our thermostat follows the following protocol:*

state	$-$	$+$
action	*turn-on*	*shut-off*

2.2 THE AGENCY HYPOTHESIS

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

Definition 4 *A* **belief assignment** *is a function,* $B : L_{\mathcal{A}} \rightarrow 2^S$*, such that for all* l *:* $B(l) \neq \emptyset$ *and* $B(l) \subseteq PW(l)$*.*

Example 1 *(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.,* $B(+) = \{hot\}$*.*

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 remark, that (after adding interpretations to each world) this approach yields a $KD45$ belief operator.[3]

However, our view is that belief really makes sense as part of a fuller description of the agent's mental level. In order to describe this mental level and to relate it to the agent's behavior, additional notions are required. We start with the agent's preference order over the set of possible states, represented by a *utility function*. This preference order embodies the agent's desires.

Definition 5 *A* **utility** *function is a function* $u : S \rightarrow \mathbb{R}$*.*

It is well known ([von Neumann and Morgenstern, 1944]) that a utility function can represent preference orders satisfying certain assumptions, which in this paper we will accept. This means that for any two states s_1, s_2: s_1 is preferred over s_2 *iff* $u(s_1) > u(s_2)$.

Example 1 *(continued): The goal of our thermostat is for the room temperature to be* ok*. This can be represented by a utility 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*.*

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). We can therefore evaluate a protocol by looking at the utility of the state it would generate at the actual world. However, due to uncertainty about the state of the world, the agent considers a number of states to be possible. It can then subjectively assess $\mathcal{P}$ in a local state l by a vector whose elements are the utilities of the plausible states $\mathcal{P}$ generates, i.e., the worlds generated by using $\mathcal{P}$ at $B(l)$.

[3] Incidentally, this gives a relation between knowledge and belief similar to the one proposed by Kraus and Lehmann in [Kraus and Lehmann, 1988].

Definition 6 *Given a context C and a belief assignment, B, with an arbitrary, fixed, order on the set $B(l)$, for every l; the* **perceived outcome** *of a protocol $\mathcal{P}$ in l is a tuple whose kth element is the utility of the state generated by applying $\mathcal{P}$ in C, starting from the kth state of $B(l)$.* [4]

Example 1 *(continued): We can construct the following table for the thermostats possible actions:*

	cold	*ok*	*hot*
turn-on	*1*	*0*	*0*
shut-off	*0*	*1*	*1*

If the thermostat 'knew' the precise state of the world, e.g., that it is cold, *it would have no trouble choosing the action* turn-on *as most preferred. When there is uncertainty, e.g., $B(l) = \{\text{cold}, \text{ok}\}$, the thermostat associates a perceived outcome of $(1, 0)$ with the action* turn-on, *and a perceived outcome of $(0, 1)$ with the action* shut-off.

While utilities are easily compared, it is not a-priori clear how to compare perceived outcomes, thus, how to choose among protocols. A strategy for choice under uncertainty is required, 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 perceived outcomes, returning the *set* of most preferred among them.

Definition 7 *A* **decision criterion** *is a function $\rho : \bigcup_{n\in\mathbb{N}} 2^{\mathbb{R}^n} \rightarrow \bigcup_{n\in\mathbb{N}} 2^{\mathbb{R}^n}$ (i.e. from/to sets of equal length tuples of reals), such that for all $\mathcal{U} \in \bigcup_{n\in\mathbb{N}} 2^{\mathbb{R}^n}$ $\rho(\mathcal{U}) \subseteq \mathcal{U}$.*

Two decision criteria we have encountered are *maximin*, which chooses the tuples in which the worst case outcome is maximal, and the *principle of indifference* which prefers tuples whose average outcome is maximal[5] (A fuller discussion of decision criteria appears in [Luce and Raiffa, 1957, Brafman and Tennenholtz, 1994]).

Returning to the example of Section 1, if Alice considers two worlds plausible, *rainy* and $\neg$*rainy*, at this order, the perceived outcome of the action *take umbrella* is $(5, -1)$, while the perceived outcome of *leave umbrella* is $(-4, 10)$. If Alice uses *maximin* she prefers $(5, -1)$, with a worst case outcome of -1, over $(-4, 10)$, with a worst case outcome of -4. She will therefore take the umbrella. Under the *principle of indifference*, Alice prefers $(-4, 10)$, with an average utility of 3, over $(5, -1)$, with an average utility of 2, and will leave the umbrella. Notice how the perceived outcome depends on Alice's beliefs. Had Alice believed only $\neg$*rainy* to be plausible, the perceived outcome of *take umbrella* would be a singleton, (-1).

We remark that decision criteria such as *maximin* can be employed with preference relations satisfying assumptions weaker than those of [von Neumann and Morgenstern, 1944].

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

Definition 8 *The* **agency hypothesis***: the agent follows a protocol whose perceived outcome is most preferred (according to the agent's decision criterion) among the set of perceived outcomes of all possible protocols.*[6]

The agency hypothesis takes the view of a rational balance among the agent's beliefs, utilities, decision criterion and behavior. It states that the agent chooses actions whose perceived outcome is maximal according to its decision criterion. Thus, the choice of the protocol is dependent upon $B(l)$ and u, which define the perceived outcome, and ρ, which helps choose among the different protocols, based on their perceived outcome. The agency hypothesis states that these components are related via this 'rationality' constraint.

3 ASCRIBING BELIEF

We now show how belief can be ascribed according to our framework. We will assume that we are ascribing a complete belief assignment to an agent, i.e., one that is defined in all local states. In many applications one can only ascribe partial belief assignments, e.g., if observations of the agent's actions exist only in some states. It is quite straightforward to generalize our discussion to this case.

Belief can be ascribed once we have certain information regarding the agent. We see this information as putting the agent in some (extended) context, which specifies some of the elements of the rational balance we have just discussed. Our strategy is to look for belief assignments confirming the agency hypothesis. That is, beliefs that would lead an entity satisfying the agency hypothesis to act according to the given protocol when its utilities and decision criterion are as given. This is a process of constraint satisfaction, where our belief assignment is constrained by the given extended context.

Definition 9 *An* **extended context** *is a 3-tuple,*

[4] For simplicity we assume a finite number of states. In the general case we use functions instead of tuples, eliminating the need to order $B(l)$.

[5] With an infinite set of tuples, *maximin* and the *principle of indifference* may not have a set of most preferred tuples. This is fixed by, for example, choosing some cutoff point.

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

$\mathcal{C} = \langle \tau, u, \rho \rangle$ *(where, τ, u and ρ are as previously defined). Given an extended context $\mathcal{C}$, a belief assignment B is* **consistent with $\mathcal{A}$'s protocol**, $\mathcal{P}_{\mathcal{A}}$, *if it confirms the agency hypothesis regarding $\mathcal{A}$.*

It is clear that this approach could be used to assign other mental states that are part of the agency hypothesis, e.g., we can ascribe goals (i.e., utilities) based on the agent's beliefs, decision criterion, and actions. We have chosen to concentrate on belief assignment. (This choice is discussed in Section 10.) The problem of belief ascription can now be formally stated as:

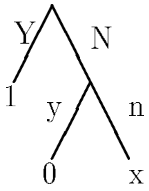
> In an extended context $\mathcal{C}$, what belief assignments are consistent with the agent's protocol, if any?

Once we define an agent's beliefs as those belief assignments that make it satisfy the agency hypothesis, we obtain an interesting characterization of belief, which stems from the grounding of beliefs in actions. The agency hypothesis implicitly defines the states an agent believes in as precisely those states that affect its choice of action.

To better understand this, we must consider an hypothetical case. Our framework attempts to do away with some of the unnatural assumptions of other approaches (see the discussion in Section 10.2). However, in principle, one can examine the case where we are supplied with information regarding how an agent would act in some local state, under any set of possible actions. In that case, we can determine whether some state s is in $B(l)$, by asking whether there are two actions, a_1 and a_2, such that both actions have precisely the same outcome on all states in $PW(l) \setminus \{s\}$, but differ on s. Roughly speaking, if one action is strictly preferred to the other (i.e., would always be chosen), we can conclude that $s \in B(l)$.[7] Of course, in practice, one has much less information, so that the choice of action can only constrain the beliefs. Still, this approach of defining belief as what affects one's choice of action is built into the agency hypothesis. This definition of belief is closely related to the definition of null-states in [Savage, 1972], as well as the definition of belief in [Morris, 1994].

Example 1 *(continued): Given our knowledge of the thermostat, what beliefs can we assign it? We know the thermostat's protocol and goals. We will assume that its decision criterion simply prefers tuples that are not dominated by another tuple. 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 perceived outcome of the action prescribed by its protocol would be strictly less preferred than the perceived outcome of the other action.*

Example 2 A simple game *The following tree describes a one-person decision problem based on a game that appears in [Kreps and Wilson, 1982]:*



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 simple question: what can we say if we observed the agent's first move to be N? This is an interesting question 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 only n to be plausible. If $x \geq 2$ then a belief that both worlds are plausible would also cause N to be preferred.*

Another decision criterion is minmax 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 minmax regret, *if $x < 2$ the*

[7] To make this precise, we would also need to have some flexibility in the values of the outcomes in the different states. But this is still in line with our explanation.

agent must believe n to follow N, otherwise it may believe either n or $\{n, y\}$.

4 CHOOSING AMONG BELIEF ASSIGNMENTS

As we observed in the thermostat example, there are often more than one consistent belief assignment. This is not surprising, as we often require additional assumptions to ascribe unique beliefs to agents, or we may need some lower level, implementation dependent, information. Dennett [Dennett, 1987] paraphrases the Duhemian thesis in this area, saying that belief and desire attribution are under-determined by the available data.

Indeed, one way of obtaining a unique belief assignment in the thermostat example would be to use a better model. That is, by using domain specific information. Assume, for instance, that the thermostat prefers not to change the course of action it is pursuing, if the result is not expected to improve its utility, i.e., if currently it is supplying hot water to the radiator then, all other things being equal, it prefers not to change this and shut-off the water supply. This assumption can be incorporated into our model by adding the course of action pursued into the state description and appropriately changing the utility function to reflect the above consideration. In that case we may be able to limit the number of consistent belief assignments

However, there are also domain *independent* assumptions and preferences that we can make when ascribing beliefs. These assumptions narrow down our choice, without changing the model used. We look at two such assumptions.

A common bias is to favor models that offer adequate explanation of the data. This is the idea behind the following:

Definition 10 *A consistent belief assignment is* **choice complete** *(within an extended context) if for all local states, the decision criterion returns a unique perceived outcome.*

Assume that in all local states no two protocols have the exact same perceived outcome. In that case, given a consistent choice complete belief assignment, no protocol is as preferred as the actual protocol. Thus, the agent will not be indifferent among a number of most preferred protocols. In this sense, a choice complete belief assignment fully explains/justifies the agent's choice of action.

Example 1 *(continued): We have seen that any belief assignment for state $-$ that includes the state* cold *is consistent. There are 4 such possibilities. However, only one of them, $B(-) = \{cold\}$ is choice complete. Given this belief assignment the agent* must *choose the action* turn-on*, while given any of the other 3 belief assignments, the agent is indifferent to the choice between* turn-on *and* shut-off*.*

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

Definition 11 *A belief assignment B is* **more general** *than B' if $\forall l \in L_{\mathcal{A}} : B'(l) \subseteq B(l)$ and $B \neq B'$. Given a set of belief assignments, $\mathcal{B}$, $B \in \mathcal{B}$ is a most general belief assignment (***mgb***) w.r.t. $\mathcal{B}$ if there is no $B' \in \mathcal{B}$ such that B' is more general than B.*

Example 1 *(continued): Any belief assignment that is a non-empty subset of $\{ok, hot\}$ is choice complete for the state $+$. However, the most general choice complete belief assignment for that state is precisely $\{ok, hot\}$.*

In the sequel we will usually assume that either the generality bias is accepted or the combination of both which prefers the most general in the set of consistent, choice complete, belief assignments. As the following lemma shows, in some sense, the latter is the best we can do in terms of assigning beliefs that do not make the agent's actions arbitrary.

Lemma 1 *If B is most general choice complete, the decision criterion satisfies the* sure-thing *principle*[8]*, and in local state l two protocols have the same perceived outcome, then there is no choice complete belief assignment under which their perceived outcome in l differs.*

Example 1 *(continued): To summerize, we have the following unique most general choice complete belief assignment for the thermostat:*

state	$-$	$+$
belief	*cold*	*not-cold*

5 ADDING TIME

Because we assumed that the thermostat has no memory nor that the environment has some special dynamics, we were able to model them without explicitly introducing time. However, time is essential for reasonably modelling many situation. Indeed the added dimension of time allows us to examine the way the mental state of an agent changes as it obtains new information.

[8] That is, if it chooses v out of $\{v, u\}$ then it chooses $v \circ w$ out of $\{v \circ w, u \circ w\}$, where $\circ$ is the concatenation operator.

We incorporate time by adding the notion of a *run*, a description of a full history of the system, and the notion of an *initial global state*, a state from which the system can start out.

Definition 12 *Let $\mathcal{G}_0 \subseteq L_{\mathcal{A}} \times L_{\mathcal{E}}$ be the set of* **initial (global) states**. *A* **run** *is a sequence of states $s_0, s_1, \ldots$ such that $s_i \in L_{\mathcal{A}} \times L_{\mathcal{E}}$, $s_0 \in \mathcal{G}_0$ and $(\forall k > 0)$ $(\exists a \in A_A) : \tau(s_{k-1}, a) = s_k$.*[9] *The* **extended system**, *$\mathcal{R}$, is the set of all possible runs.*

Having changed from static states to runs, we must redefined some of our basic notions.

Definition 13 *The set of* **possible worlds**, *$S = \{s | s$ is a global state appearing in a run in $\mathcal{R}$ $\}$. A* **context** *is redefined as $C = \langle \tau, \mathcal{G}_0 \rangle$ and an* **extended context** *is redefined as $\mathcal{C} = \langle \tau, \mathcal{G}_0, u, \rho \rangle$. We redefine the utility function as $u : \mathcal{R} \rightarrow \mathbb{R}$.*

Applying a protocol $\mathcal{P}$ at a state s will generate a unique run r whose initial state is s, where each state of r is obtained by performing the action prescribed by $\mathcal{P}$ at the previous state. This allows us to maintain the notion of a perceived outcome because we can now associate a utility with each protocol at each state, the utility of the run that this protocol induces at that state.[10]

One last adjustment; we defined a belief assignment as a function $B : L_{\mathcal{A}} \rightarrow 2^S$. This definition will make it hard for us to investigate belief change, i.e., the relations between an agent's beliefs at different states of a run. For example, if the agent has a clock, then its local state at two consecutive states of a run will differ, because in each the clock's value would be different; consequently, the states the agent considers plausible at these local states would be disjoint. Rather than add additional atemporal elements, such as an explicit language, we overcome this problem by redefining a belief assignment as assigning *possible runs*, rather than *possible worlds*, i.e., $B : L_{\mathcal{A}} \rightarrow \mathcal{R}$. Because runs are atemporal object, this choice makes the fundamental changes in an agent's beliefs more clearly visible. [11]

6 BELIEF CHANGE

With time added to our model, we must start considering how the agent's mental state changes over time. Belief ascription, as currently defined, allows erratic change across local states. An extreme example would be an agent whose local state changes from l to l', such that $PW(l) = PW(l')$, yet $B(l) \cap B(l') = \emptyset$. Part of our conception of agents involves an expectation that their beliefs should change in a 'sensible' way ([Alchourron *et al.*, 1985]). Constraints on belief change across states are also of immense importance if we are to be able to predict an agent's behavior. Having ascribed beliefs to the agent based on past actions we must have such constraints to deduce the agent's current beliefs. Having deduced the new beliefs, we can use them to predict the agent's choice of action.

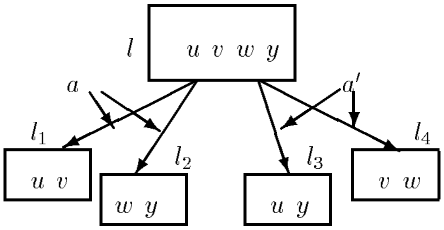


Figure 1: The change in an agent's local state after performing actions a and a', respectively.

We will look at two patterns of belief change that we find reasonable and prove two representation theorems. The theorems show that there are two ways of viewing these restrictions, either as constraints on new beliefs imposed by the previous beliefs and the new information, or as requiring a general static way of representing the agent's beliefs. We can then incorporate these restrictions into our model by requiring a belief assignment to be consistent in the static sense of Definition 9, and to exhibit the desired pattern of belief change. This will redefine the problem of belief ascription for agents that can acquire new information while acting.

In what follows we will assume that the agent has *perfect recall*, i.e., its local state contains all previous local states. This describes agents that do not forget. However, much of the following development also makes sense when the agent has only partial memory of past states. Perfect recall implies that an agent's local state changes from one state to the next. Therefore, any two states on the same run are distinguishable.

6.1 ADMISSIBILITY

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

We can illustrate this using Figure 1. The agent is initially in local state l, where the possible runs are u, v, w and y. Assume that $B(l) = \{u, w\}$. After performing action a the agent finds itself in state l_1. If the

[9] Finite runs are modelled by runs in which $\exists\, n \forall m\ s_n = s_{n+m}$.

[10] Notice the this requires extending the utility function over suffixes of runs. This is quite straightforward given our deterministic model of the environment.

[11] The interested reader may consult [Friedman and Halpern, 1994], where belief change is investigated from this perspective.

agent's beliefs are admissible then $B(l_1) = \{u\}$. However, assume that $B(l) = \{u, v\}$ 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 $B(l_2) \subseteq \{w, y\}$).

Definition 14 *A belief assignment B is* **admissible**,[12] *if for local states l, l' such that l' follows l on some run: whenever $PW(l') \cap B(l) \neq \emptyset$ then $B(l') = PW(l') \cap B(l)$; otherwise l' is called a* **revision state** *and $B(l') \subseteq PW(l')$ is otherwise not restricted.*

If we were to assume that the worlds here are models of some theory then, in syntactic terms, admissibility corresponds to conjoining the new data with the existing beliefs, whenever this is consistent. It is closely related to the probabilistic idea of conditioning our beliefs upon new information. Most work on belief revision makes additional requirements on beliefs following inconsistent information (what we call a revision state). We will return to this issue in the end of this section.

We can shed additional light on this restriction by 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. That is at each state l the set $B(l)$ is exactly the set of elements in $PW(l)$ that are minimal w.r.t. this ranking.

Definition 15 *A* **well founded ranking** *r of a set Q is a mapping from Q to a well ordered set $\mathcal{O}$. Given a subset Q' of Q, the elements minimal in Q' are those that have the minimal rank, i.e., are assigned the lowest element of $\mathcal{O}$ by r.*

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 a minimal element. The elements of lower rank are considered to be better, more preferred, or more likely.

Theorem 1 *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 runs such that $B(l) = \{s \in PW(l) : s \text{is r-minimal in } PW(l)\}$.*

6.2 WEAK ADMISSIBILITY

The requirement that belief assignments be admissible may seem too strong. A weaker requirement is the following: if my new state is consistent with a run I believed before, I should still believe in that run's possibility. However, unlike when my belief assignment is admissible, once I learn that a run I considered plausible before is in fact impossible, I may additionally consider plausible runs which I did not consider plausible before. However, if what I learn only reaffirm my previous beliefs, i.e., I only learn that a run I did *not* believe plausible is completely impossible, my beliefs should not change. Formally:

Definition 16 *A belief assignment is* **weakly admissible** *if when a local state l' follows l,*

1. *$B(l') \supseteq B(l) \cap PW(l')$.*
2. *If $B(l) \subseteq PW(l')$ then $B(l') = B(l)$*

Looking at Figure 1 again, if the agent believed in u, w in l and its state changes to l_1 then it may believe either in u or in u, v. However, if the agent only believed u to be plausible in l, then at l_1 its only consistent belief is in u.

Fortunately, we can again relate the ascription of weakly admissible beliefs to that of ascribing a static partially ordered belief structures. Again, this structure determines the agent's beliefs at l by choosing the minimal elements of $PW(l)$ according to this structure.

Definition 17 *A* **partial pre-order** *on Q is a partial subset of $Q \times Q$ that is reflexive and transitive.*

Theorem 2 *The beliefs of an agent with perfect recall are weakly admissible iff there is a partial order $<$ on the set of possible runs, such that its beliefs at l correspond to the minimal runs in $PW(l)$ according to $<$.*

Patterns of beliefs change similar to ours emerge in the work of other researches (e.g., [Friedman and Halpern, 1994, Lamarre and Shoham, 1994]). Indeed, relations between belief revision and belief update, and representations using partial and total pre-orders are well known. It was shown in [Katsuno and Mendelzon, 1991b] that any revision operator that satisfies the AGM postulates ([Alchourron *et al.*, 1985]) can be represented using a ranking of the set of possible states. We require less to obtain the same representation. The reason for this, besides our assumption of perfect recall, is our emphasis on belief ascription, rather than on prescribing belief change. The need for additional requirements arises when counter-factual reasoning has to be accounted for. Then, given a certain state, all ways in which it can be revised must be accounted for. On the other hand, we are not asking the question of how the agent's beliefs would look like if it were to take a different action than the one prescribed by its protocol; we only need to explain the particular actions performed by the agent at different states.

[12] This is not to be confused with the notion of admissibility in game theory.

7 EXISTENCE - GOAL SEEKING AGENTS

But when does a belief assignment exist? From the point of view of modelling this question is crucial, and Savage's answer to it ([Savage, 1972]), provides much of the foundation of statistics and economic modelling. In order to model programs, machines, or humans, using the various abstract mental states investigated in AI, it is important to recognize the conditions under which these modelling tools can be used.

Examining Savage's work we see that he is able to ascribe likelihood and utilities by imposing certain consistency restrictions on the agent's actions. We will follow a similar path. We first restrict ourselves to a certain class of extended contexts and then require the agent's protocol to satisfy two restrictions. We will show that an agent satisfying these restrictions and operating in the given class of extended contexts, can be ascribed a unique most general choice complete belief assignment.

The contexts we examine here are of a special kind that is quite natural in many AI applications. Local states are of two types, goal states and non-goal states. Runs are finite and their utility is determined by the last local state, i.e., 1 if it is a goal state, and 0 otherwise. We have a distinguished action, HALT, whose utility (or more precisely, that of its outcome) in a goal state is 1, and 0 otherwise.

We define two rationality postulates on protocols, that embody a notion of a *goal-seeking* agent. 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. The *rational despair* postulate says that to halt the agent must either be in the goal or be able to show a possible world under which he can never reach the goal. Notice that these postulates refer to the set $PW(l)$ describing the agent's knowledge, rather than to $B(l)$ (preventing possible circularity later).

Rational Effort Postulate The protocol in a local state l is either HALT or weakly dominates HALT.

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

We will call an agent satisfying these postulates who operates in the contexts described above and whose decision criterion is consistent with weak dominance (i.e., if v is preferred over v' then v' does not weakly dominate v[13]), a *goal-seeking* agent.

Theorem 3 *If A is a goal seeking agent then it can be ascribed a unique most general admissible belief assignment and a unique most general choice complete admissible*[14] *belief assignment.*

Many people view rational choice as equivalent to expected utility maximization under some probability distribution. While we find the probabilistic approach most appropriate in many contexts, we do not share this view (see the following discussion). Indeed, we show that, in 0/1 utility 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 (subjective) probability assignment, whose preferences are represented by a 0/1 utility function, and whose decision criterion is based on expected utility maximization w.r.t these probability and utility assignments. However, we require that when no action has an expected utility greater than 0 then HALT is performed.

Corollary 1 *An agent that can be modelled as a B-type agent is a goal-seeking agent, and consequently, can be viewed as a perceived outcome maximizer, using some admissible belief assignment and a decision criterion consistent with weak dominance.*

8 COMPLEXITY - PRACTICAL AGENTS

In the following theorems we will assume specific (though very general) representation of the protocol and the context. We defer these details to the end of this section.

Theorem 4 *The problem of finding an mgb for the goal cardinality criterion is CO-NP hard, where the input consists of a description of the context and the protocol as defined below.*

This result shows that even for a simple commonsensical criterion, computing an mgb may be very expensive. A closer examination of its proof suggests that a more practical view of belief may greatly help simplify the problem. This may be described as the wishful-thinking approach. The states I believe in are those I consider plausible *and* from which I think I have some chance of achieving the goal. Such beliefs may be interpreted as not simply plausible worlds, but rather *practical* plausible worlds, i.e. plausible world in which I can do something to achieve my goal. For example, in the game of bridge this is a common tactic for the defenders. If it seems likely that the contract can be made, the defenders play under the assumption that they can actually bring down the contract, if this is possible.

[13] Let $v(i)$ be the ith element of v. We say that v' *weakly dominates* v if $\forall i\ v'(i) \geq v(i)$ and $\exists i\ v'(i) > v(i)$.

[14] The notion of generality as defined in Section 3 must be adjusted in the case of admissible belief assignments. See Section 8.

Practical beliefs can be captured by the following constraint: $\neg K \neg \Diamond g \rightarrow B \Diamond g$ (where $\Diamond g$ is satisfied in a world in a run if the run eventually satisfies g). For agents that can be consistently ascribed such beliefs the problem of belief ascription becomes tractable.

Theorem 5 *If the agent can be ascribed beliefs that obey the following condition*

$$\neg K \neg \Diamond g \rightarrow B \Diamond g$$

then the problem of finding the mgb given the goal cardinality criterion becomes polynomial in the input size $\times$ *the maximal length of a run.*

The representation The agent's state is described by two components: his physical state and his memory. His protocol is a function of both. Both the agent and the environment have a finite number of states and the transition function is described by a finite state machine. The finite state machine consists of two components (i.e., it is a product of two finite state machines), one of them being the physical description of the agent, the other corresponding to the environment. The actions of the agent and the environment together determine the next state of both (i.e. the physical state of the agent and the environment). Thus, this models the common situation in which the result of the agent's actions depend on his physical state and on the state of the environment. The protocol may depend on the agent's history as well (i.e., not only on its physical state) and will be represented as a decision tree of polynomial depth.

A more precise statement of the complexity in Theorem 5 is that it is linear in the number of initial states $\times$ the maximal length of a run $\times C$, where C is the time it takes to compute the next state given the current state, and is bounded by the size of the protocol $\times$ the size of the transition function.

9 CONDITIONS FOR A UNIQUE BELIEF ASSIGNMENT

In Section 4 we looked at two criteria for choosing among consistent belief assignments, in particular the notions of generality and a most general belief assignment (mgb) were introduced. Naturally, it is desirable to be able to point out a single consistent belief assignment as most appropriate. If one accepts the generality criterion, this translates into the question of whether a unique most general belief assignment exists. In general the answer is no and we provide a counter example. However, after adopting the notion of generality to fit the case of admissible and weakly admissible belief assignments, we will show that for monotone decision criteria we can give a positive answer to the uniqueness question. All of our results can be extended to the case of unique most general choice complete belief assignments.

The question of uniqueness is closely related to a property of decision criteria that we now discuss. Remember that the tuples on which we are making a decision correspond to the perceived outcomes of a protocol in some local state. An interesting question is the following: If $\mathcal{P}$ is preferred over all other protocols when the plausible worlds in $PW(l)$ are $B_1(l)$ and also when the plausible worlds are $B_2(l)$; is it still preferred when the plausible worlds are $B_1(l) \cup B_2(l)$? When $B_1(l)$ and $B_2(l)$ are disjoint, a positive answer is reminiscent of the *sure thing principle*, and is highly intuitive. We call a decision criterion satisfying this property *weakly monotone*. When $B_1(l)$ and $B_2(l)$ are not disjoint, if this property is satisfied, we say that the decision criterion is *monotone*. The *maximin* and *minmax regret* criteria satisfy this property, but the *principle of indifference*, for instance, does not.

Theorem 6 *If the decision criteria is monotone and a belief assignment exists then there is a unique mgb.*

It is easy to construct a counterexample when monotonicity is not obeyed, using the principle of indifference.

Example 3 *We are presented with the following decision problem.*

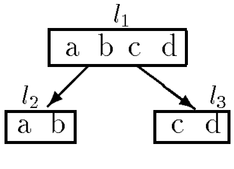
	s_1	s_2	s_3	s_4	s_5
a_1	*2*	*2*	*11*	*2*	*2*
a_2	*7*	*7*	*0*	*7*	*7*

If we believe in $\{s_1, s_2, s_3\}$ *or* $\{s_3, s_4, s_5\}$ *then according to the principle of insufficient reason we prefer to take action* a_1*. However, for there to be a unique mgb their union,* $\{s_1, s_2, s_3, s_4, s_5\}$*, must be consistent with* a_1*, yet given this set of beliefs we prefer action* a_2*.*

9.1 UNIQUE ADMISSIBLE BELIEFS

When an agent has admissible beliefs, the notion of a most general belief makes little sense. Because of its properties, we will usually have to make the belief assignment less general in one state in order to obtain one that is more general in another state

Example 4 *Consider the following case: there are four possible runs:* a, b, c, d *in the initial local state* l_1*. After the first action there are two possible local states* l_2 *and* l_3 *corresponding to two sets of two runs each:* a, b *and* c, d*.*



Assume that in l_1 *we can ascribe beliefs in* a, b *or* a, b, c*, in* l_2 *we can ascribe belief in* a, b*, and in* l_3 *we can* *ascribe belief in c or c, d. There are two admissible belief assignments: B_1 assigns a, b, c to $s1$, 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.*

However, the notion of a ranking helps us recognize a preference criteria over admissible beliefs. In another context, that of non-monotonic logics, worlds minimal in a ranking structure are often described as 'most normal'. There, structures that are 'thicker' at the bottom are often preferred, because they make less assumptions of non-normality. Our preference for such structures is that the actions one takes initially, when there is least knowledge, are the most crucial, thus the assumptions made at the initial states, which correspond to the minimally ranked states, are the most crucial. We prefer belief assignments that take more worlds into consideration in these stages. Such beliefs are less likely to be inconsistent.

Definition 18 *An admissible beliefs assignment B is more general than B' if, represented as ranking functions, B and B' are identical up to some rank m, and $B^m \supset B'^m$ (recall that B^m is the m^{th} rank of B).*

We will refer to this definition of more general when talking about mgb in the context of admissible beliefs.

Theorem 7 *For agents with perfect recall, if the decision criteria is monotone then the set of consistent admissible belief assignments, if non-empty, contains a unique mgb.*

9.2 UNIQUE WEAKLY ADMISSIBLE BELIEFS

For admissible belief assignments the property of set-theoretic inclusion for more general belief assignments had to be given up. This was unfortunate because the definition of generality based on set-theoretic inclusion guarantees that beliefs attributed to the agent according to the mgb can be attributed given any consistent belief assignment. Fortunately, for weakly admissible belief assignments we can see, via their representation as partial orders, that a most general belief assignment in the set theoretic sense, exists. This is pleasing because it would otherwise be difficult to motivate a definition of one partial order being more general than another in which all partial orders were comparable.

Theorem 8 *For monotonic decision criteria if the set of belief assignments is not empty then a (set theoretic) most general weakly admissible consistent belief assignment exists.*

10 DISCUSSION

To conclude we re-examine the work presented in this paper and some related research.

10.1 RE-EXAMINING THE FRAMEWORK

The ability to model agents at the mental level is most likely required for any form of artificial intelligence. However, as an abstraction it is already useful for more mundane modelling tasks. It is extremely important in multi-agent domains, as agents must construct models of other agents, but it is also useful as a means of describing and analyzing systems at an abstract, yet highly intuitive, level. As such, a model of the mental level should strive to be simple and intuitive. Yet, it must also be precisely formulated with sound foundations. We believe that the framework we presented meets these criteria. In its formulation we aspired for a framework that will shed light on the semantical and theoretical issues of belief ascription without making some of the problematic modelling assumptions made in other frameworks (discussed next), that make their use in practical applications difficult.

Our work attempts to improve our understanding of belief ascription on two levels, as a practical modelling task and as a way of grounding the notion of belief in an agent's choice of action through an ascribed decision making process. Belief in our framework, represents the agent's subjective information on the outside world that is utilized in decision making. Thus, we modelled beliefs as a function of the agent's local state, for otherwise, the actual state of the world would affect its beliefs, without affecting its state. We demonstrated that for a large class of goal-seeking agents, beliefs can be ascribed within our framework. We have also suggested two methods for narrowing the choice among candidate belief assignments that under certain conditions point out a unique belief assignment and investigated the computational feasibility of belief ascription. While this is a difficult task in the general case, we showed that for certain types of agents, this task is tractable.

One may ask why do we emphasize belief ascription, when the framework supplies the basis for ascribing utilities or a decision criterion. Ascription of these notions is certainly important, but there are a number of reasons for our choice. Belief and knowledge are by far the most extensively researched mental states within AI and philosophy (e.g. [Kripke, 1963, Hintikka, 1962, Moses and Shoham, 1993, Katsuno and Mendelzon, 1991a, Alchourron *et al.*, 1985, Goldzmidt and Pearl, 1992, Boutilier, 1992, del Val and Shoham, 1993, Lamarre and Shoham, 1994, Friedman and Halpern, 1994]), and it is therefore important to understand where they come from and how to ascribe them. Moreover, mental-level modelling is often used by us to construct rough descriptive models. It is often the case that an agent's goals are known. This suffices to supply rough estimates of utilities. Knowing an agent's decision criteria seems harder, but we have shown that for 'reasonable' protocols in 0/1 utility contexts, beliefs can be ascribed based on the trivial assumption that the agent prefers weakly dominant tuples. These contexts are natural in many CS applications. Additionally, while the plausible worlds for an agent in different situations may be unrelated, the decision criterion is almost constant. Observing an agent's decision criterion in one case seems a good indicator of its decision criterion in other cases. Naturally, in normative applications, such as analysis of protocols, the designer can readily provide all the required information.

10.2 RELATED WORK

There has been some important research on ascribing mental states to agents. One major research area is plan ascription, an important task in discourse understanding and multi-agent systems (e.g.,[Kautz, 1990, Konolige and Pollack, 1989, Pollack, 1990]). The aims of the work on plan ascription is more specific than ours and plans are often ascribed based on utterances (e.g., [Konolige and Pollack, 1989]). More specifically, Konolige ([Konolige, 1990]) has done some theoretical work on explanatory belief ascription. His work looks at the question of how to explain known beliefs of an agent by ascribing this agent additional beliefs. His work implicitly assumes a high-level agent into whose beliefs we have some access, usually through the utterances of that agent. He then explains these beliefs based on other, ascribed, beliefs. This work does not deal with the general problem of belief ascription. Both [Konolige, 1990, Konolige and Pollack, 1989] have a somewhat syntactic flavor, due to the use of argumentation systems and derivational models. In contrast, our framework does not employ some of the stronger techniques used in these papers, but addresses more basic semantic issues and is based upon a more general semantic model of the mental level. It seems general enough to provide the foundations for belief ascription based on utterances, if utterances are treated as speech acts [Austin, 1962].

Most influential on our work was the knowledge ascription framework of [Halpern and Moses, 1990] and the closely related work of [Rosenschein, 1985] on situated automata, which defines knowledge in a similar manner. Halpern and Moses define a formal notion of knowledge that is grounded in the state space description of a system. The set of possible worlds of an agent then emerges from the notion of a local state. We have built our framework upon their framework. However, the notion of knowledge does not take into account the agent's actions, but is based strictly on the state-space description of the domain. By adding action into the picture, we are able to enrich the model.

The work of Savage [Savage, 1972] and Anscombe and Aumann ([Anscombe and Aumann, 1963]) on subjective probability and choice theory is closely related to our work. This work takes a similar view of the notion of belief, i.e., as emerging from the representation of an agent's preference over actions. It shows that given the preferences of an agent over possible actions, where these preferences satisfy certain constraints, the agent can be viewed as acting as an expected utility maximizer under unique ascribed (probabilistic) beliefs and utilities. This work is extremely elegant, yet it has some limitations from our perspective. First, it is probabilistic, while much work on knowledge and belief within AI, CS, and philosophy is discrete. This means that it cannot provide the foundation for these notions of belief. Secondly, there are serious practical problems with its application to our setting. The strength of the representation theorems of Savage and Anscombe and Aumann, especially the uniqueness property, stems from a number of strong requirements. One must supply a total pre-order on the set of all possible acts, i.e., *all* functions from initial states to outcomes, many of which are purely fictional acts. This information will not be available to an observer of the system, nor will it be easy for a designer to come up with it. Another assumption is that the state description is rich, i.e., that for any natural n, there exists a partition of the set of states into n subsets, all of which are equally likely. This means that the number of states must be infinite. In addition, expected utility maximization has been criticized as a normatively inadequate decision criterion (see [Kyburg, 1988] and other papers in [Gärdenfors and Sahlin, 1988]). It is certainly inadequate descriptively, as many studies have shown (e.g., [Machina, 1989]) and is therefore problematic in modelling other agents.

In contrast, we believe that our formalism is better suited for many modelling tasks in AI and CS. Our framework is discrete, thus relevant to the body of work on discrete notions of belief. It also requires much less information. We only assume awareness of the agent's actual protocol (e.g., through observations or as a given specification), and knowledge of the possible alternatives, that is, the agent's set of possible actions. It does not require complete knowledge of the preference relation among other protocols, nor the additional richness assumption on the set of states. Moreover, as we remarked earlier, it is straightforward to apply our ideas when only part of the protocol is known, for example, when we have seen the agent act only in a subset of its set of possible local states.

Our greater generality cannot come without a price. Because we require less information, in the general case, we cannot supply a unique belief assignment, but only constrain the set of consistent ones. Indeed, we must also require information on the agent's preference over specific outcomes (which in this paper took the form of a utility function)[15] and the decision criterion (although we saw one example where the latter was not necessary).

Another advantage of our framework is that it leaves the decision criterion as a parameter. This gives us added modelling flexibility. On the one hand, our notion of a decision criterion can be easily generalized for our framework to cover expected utility maximization as a special case. On the other hand, decision criteria that allow for notions of preference that do not satisfy the Von Neumann-Morgenstern axioms are possible. This flexibility is useful for descriptive purposes, because we may want to model different classes of entities. But even for normative purposes one may wish to relax these requirements. For example, we may want our agent to act as if some goal's utility is infinitely greater than any other, e.g., the preservation of human life. This can only be done if we relax the Von Neumann-Morgenstern axioms.

We have stressed the problem of grounding before. Research on abstract mental states is certainly important. Yet, notions such as 'beliefs', 'goals','intentions', etc., are more meaningful if they can be embedded in some concrete setting. A major contribution of the work of [Halpern and Moses, 1990, Rosenschein, 1985] is supplying this concrete setting, showing how the notion of knowledge arises in distributed systems and in situated agents. In his 1985 Computers and Thoughts speech [Levesque, 1986b], Levesque spoke about making believers out of computers, thus supplying a concrete interpretation of belief. However, the actions of the systems Levesque was referring to, all have to do with answering queries. Levesque's view serves as a means of abstracting the constraint that (for a meaningful investigation of knowledge representation schemes) the system's actions must depend on the content of the data-structures used to represent knowledge (see also [Levesque, 1986a, p. 258]). This abstract view has proven to be extremely fruitful for understanding the task of knowledge representation. However, most systems (computer, mechanical or biological) are situated in some environment. Their goal is usually much more general than correctly representing it, although that may be useful. Their actions range from writing into files to changing the temperature of a room. Therefore, there is much to be gained by taking actions in general as the basis for ascribing beliefs. If we do not de-contextualize belief by ignoring the agent's actions, goals, etc., we may be able to obtain a better understanding of these systems. Indeed, we see our main conceptual contribution in grounding the discrete notions of beliefs used in AI in the more concrete (and empirically testable) notion of preference over action. In modelling a system, ascribing it beliefs makes sense if and only if the system is *acting as though* these are its beliefs.

Acknowledgement

We would like to thank Yoav Shoham, Nir Friedman, Daphne Koller, Joe Halpern and Alvaro del Val for their help. Their comments and criticism have greatly helped us improve this work and its presentation. We also wish to thank Robert Aumann and David Kreps who were most helpful in putting this work in perspective.

This work was supported in part by grants from the National Science Foundations, Advanced Research Projects Agency and the Air Force Office of Scientific Research.

References

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

Anscombe, F. J. and Aumann, R. J. 1963. A definition of subjective probability. *Annals of Mathematical Statistics* 34:199–205.

Austin, J. L. 1962. *How to do things with words.* Oxford University Press.

Boutilier, C. 1992. Normative, subjective and autoepistemic defaults: Adopting the ramsey test. In *Principles of Knowledge Representation and Reasoning: Proc. Third Intl. Conf. (KR '92).*

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

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

Dennett, D. C. 1987. *The Intensional Stance.* MIT Press, Cambridge, Mass.

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

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

Gärdenfors, P. and Sahlin, N. E., editors 1988. *Decision, Probability, and Utility.* Cambridge University Press, New York.

Goldzmidt, M. and Pearl, J. 1992. Rank-based systems: A simple approach to belief revision, belief update and reasoning about evidence and actions. In

[15] In Savage's framework, this is simply one aspect of the ordering on acts.

Principles of Knowledge Representation and Reasoning: Proc. Third Intl. Conf. (KR '92). 661–672.

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

Hintikka, J. 1962. *Knowledge and Belief.* Cornell University Press, Ithaca, NY.

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

Katsuno, H. and Mendelzon, A. O. 1991b. Propositional knowledge base revision and minimal change. *Artificial Intelligence* 52(3).

Kautz, H. 1990. A circumscriptive theory of plan recognition. In Cohen, P. R; Morgan, J.; and Pollack, M. E., editors 1990, *Intentions in Communication*, Cambridge, Mass. MIT Press. 105–133.

Konolige, K. and Pollack, M. E. 1989. Ascribing plans to agents. In *Proc. Eleventh Intl. Joint Conf. on Artificial Intelligence (IJCAI '89).* 924–930.

Konolige, K. 1990. Explanatory belief ascription. In *Proc. of the Third Conf. on Theoretical Aspects of Reasoning About Knowledge*, San Francisco, California. Morgan Kaufmann. 57–72.

Kraus, S. and Lehmann, D. J. 1988. Knowledge, belief and time. *Theoretical Computer Science* 58:155–174.

Kreps, D. M. and Wilson, R. 1982. Sequential equilibria. *Econometrica* 50(4):863–894.

Kripke, S. 1963. Semantical considerations of modal logic. *Zeitschrift fur Mathematische Logik und Grundlagen der Mathematik* 9:67–96.

Kyburg, H. E. 1988. Bets and beliefs. In Gärdenfors, P. and Sahlin, N. E., editors 1988, *Decision, Probability, and Utility.* Cambridge University Press, New York. chapter 6.

Lamarre, P. and Shoham, Yoav 1994. Knowledge, certainty, belief and conditionalization. In *Proc. of Fourth Intl. Conf. on Principles of Knowledge Representation and Reasoning.*

Levesque, H. J. 1986a. Knowledge representation and reasoning. *An. Rev. Comput. Sci.* 1:255–287.

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

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

Machina, M. 1989. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature* 27:1622–1668.

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

Morris, S. 1994. Revising knowledge: A hierarchical approach. In *Proc. of the Fifth Conf. on Theoretical Aspects of Reasoning About Knowledge*, San Francisco, California. Morgan Kaufmann.

Moses, Y. and Shoham, Y. 1993. Belief as defeasible knowledge. *Artificial Intelligence* 64(2):299–322.

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

Pollack, M. E. 1990. Plans as complex mental attitudes. In Cohen, P. R.; Morgan, J.; and Pollack, M. E., editors 1990, *Intentions in Communication*, Cambridge, Massachusetts. MIT Press. 77–104.

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

Safra, S. and Tennenholtz, M. 1993. On planning while learning. submitted.

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

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

A Proofs

Proof of Theorem 1

Perfect recall implies that worlds that were impossible cannot become possible. For one direction note that if the agent's belief are the accessible worlds of minimal rank then when we move into a new state the possible worlds are a subset of the previous possible worlds. Since the ranking is fixed, if any of the previously minimal worlds are possible then they remain minimal, and no world that was not minimal can become minimal. If none of the previous minimal worlds is possible, then the new minimal worlds are not related in any way to the previous ones. For the other direction, assume that B is admissible, we construct a ranking function based on B. For simplicity we assume the initial states are all indistinguishable. We give an algorithm for constructing the ranking. It may require an infinite number of steps to be completed. However, each step produces a ranking that contains the previous one, i.e., all states that are part of the previous ranking have the same rank, so the answer is the union of running this algorithm for any finite amout of time. At stage n of the construction we consider the set of local states that appear at time $n-1$ of any possible run. First, we assign all the states of $B(l^i)$ to rank 0, where l^i is the initial local state and rank 0 is the lowest. At the nth stage we look at all local states l that are at stage n of some possible run and for every state. For any such state l, if $u \in B(l)$ but has not yet been assigned a rank, we assign it to rank $n-1$. We have to show that this reproduces the original belief assignment. It clearly does at l^i. Assume that it does at l, a local state that is nth in some run and let l' be a child of l. If $B(l')$ contains a run that is not in $B(l)$ then by admissibility $B(l) \cap PW(l') = \emptyset$. Therefore, non of $B(l')$ appear at a rank smaller than n, but all apear in rank n. We must also show that no run in $PW(l') \setminus B(l')$ appears in rank n. But if this was the case then in some local state appearing at stage n or before, this state was believed. Clearly, it could not have been in one of the ancestors of l', because that would violate admissibility. Therefore, it must have appeared in a different tree, but this tree would contain completely different possible runs. ▌

Proof of Theorem 2
It is clear that beliefs based on a partial ordering of states are weakly admissible. If s is minimal in a set A, it is minimal in any subset of A containing s, and the minimal items in a set A are the same as the minimal items in any subset of A that contains all these minimal items. Assume that B is weakly admissible, we will construct a partial ordering $<$ such that for each local state l our beliefs in l will be equal to the minimal elements of $PW(l)$ according to $<$. Let l_i be the initial local state. We make all states of $B(l_i)$ incomparable and minimal. Let l be a local state following l_i (remember, we are assuming full history), $PW(l) \subseteq PW(l_i)$. If $B(l_i) \subseteq PW(l)$ we change nothing. If $B(l_i) \not\subseteq PW(l)$ we choose some element s s.t. $s \in B(l_i)$ but $s \notin PW(l)$ and make all elements of $B(l)$ that are not in $B(l_i)$ above it in the ordering (i.e., less preferred), but incomparable to each other. This process is performed breadth first. Elements that remain unassigned in the end can be made less preferred than the last element assigned.

Lemma 2 *The belief assignment B' induced by the above partial order equals B.*

Proof: We prove by induction on the nodes of the tree that $B(l) = B'(l)$ and that for all l, $\{s : s \in B'(l)\}$ are incomparable. For the initial local state, l_i, we have constructed the partial order such that all elements of $B(l_i)$ are incomparable. Since all other worlds were placed above some world in $B(l_i)$ they are minimal as well. Let l be some local state and l' its parent. If $PW(l)$ contains $B(l')$ than we know that nothing changes in B' (because it is a partial order) and B (because it is weakly admissible). Otherwise we have made whatever worlds are in $PW(l)$ but not in $B(l)$ above some world s that is no longer possible. This makes all of them minimal, since s was incomparable to other worlds in $B(l')$, we now have that all worlds in $B(l)$ are minimal in $PW(l)$ w.r.t. B' and are incomparable. ▌

Proof of Theorem 3
We construct an admissible belief assignment B. 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 $B(l^i) = PW(l^i)$ or otherwise there is a world $s \in PW(l^i)$ such that the goal cannot be reached from s (using Rational Despair). We let $B(l^i)$ be 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 weakly dominant, we let $B(l^i) = S$. Such a set exists and is not empty, since otherwise $\mathcal{P}$ must be HALT (applying Rational Effort). By definition of admissibility, for any state, l, consistent with S (i.e., $S \cap PW(l) \neq \emptyset$) we must define $B(l) = S \cap PW(l)$, therefore, we need to see that in any state l consistent with S, $\mathcal{P}$ still weakly dominates according to $S \cap PW(l)$. Assume the contrary. This means that for some state $s \in S \cap PW(l)$, $\mathcal{P}$ 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 an action 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 $s \in PW(l)$ from which no protocol attains the goal. We let $B(l)$ the set of all such worlds. ▌

Proof of Theorem 4
We give a reduction to the complement of the problem of producing a guaranteed plan for goal achievement in [Safra and Tennenholtz, 1993]. There the problem of the existence of a guaranteed plan is shown to be NP-hard. In that problem we are given a local state state l consistent with n global states, along with a transition function τ and a set of goal states g. We are to show that there is no guaranteed plan for achieving the goal within a polynomial number of steps. We transform this into the problem of computing the most general belief assignment with the goal cardinality criterion by adding a new action $\bar{a}$ that moves us from all initial states but one, s, to a goal state, but from s causes us not to be able to reach the goal. s is part of $B(l)$ when the protocol consists of the action $\bar{a}$ iff there is no protocol that guarantees achieving the goal from l (i.e., from all possible global states consistent with l, including s). ▌

Proof of Theorem 5
This condition means that if from some of the global states consistent with the local state we can reach the goal then the states I consider possible are those from which the goal can be reached. Because of generality this will mean that the states believed are exactly those from which the goal can be reached. Computing these states is simple. We check for each local state if we can reach the goal from it using the protocol, if so we believe in it. We can do so by simply simulating the protocol on the set of initial states. This in fact gives

us the most general choice complete belief assignment obeying the above requirement. ∎

Proof of Theorem 6

Since the belief assignment has to satisfy local criteria, i.e., for any local state the actual protocol maximizes outcomes w.r.t. the decision criteria, it is easy to see that because of monotony, if B and B' are belief assignments then so is $B \cup B'$. This ensures uniqueness. ∎

Proof of Theorem 7

Given that the agents have complete histories 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. We show that we can construct an admissible belief assignment that is more general than both. Let l be a local state. W.l.g. assume that $B_1^0 \neq B_2^0$ (remember, B^i denotes the states in the ith rank of B.. We claim that an admissible belief assignment B exists such that $B^0 = B_1^0 \cup B_2^0$.

First we show that for local states l such that $PW(l) \cap (B_1^0 \cup B_2^0) \neq \emptyset$ we are consistent. If $PW(l)$ contains only elements from B_1^0 or B_2^0, then nothing changes since (because of admissibility) our beliefs there according to B must be the same as one of B_1 or B_2, and since both are consistent with the decision criteria, there is no problem. If $PW(l)$ contains states from both, our beliefs will be the union of these and due to monotonicity are still consistent. Otherwise, l must be a revision state, in which case $B(l)$ need not be affected by our choice of B^0 and we can arbitrarily choose $B(l)$ to be the same as $B_1(l)$. This gives us an admissible belief assignment that is more general than both B_1 or B_2. ∎

Proof of Theorem 8

First observe that the union of two weakly admissible belief assignments is a weakly admissible belief assignment. Let B_1 and B_2 be such assignments. Let l be a local state and l' be some child of l. We know that all members of $B_1(l)$ (resp. $B_2(l)$) that are in $PW(l')$ are in $B_1(l')$ (resp. $B_2(l')$). Thus all members of $B_1 \cup B_2(l)$ that are in $PW(l')$ are in $B_1 \cup B_2(l')$. If all members of $B_1(l)$ and $B_2(l)$ are in $PW(l')$ then $B_1(l) = B_1(l')$ and $B_2(l) = B_2(l')$, so $B_1 \cup B_2(l) = B_1 \cup B_2(l')$. Therefore $B_1 \cup B_2$ is weakly admissible.

To prove uniqueness we need to show that $B_1 \cup B_2$ is consistent with the rationality criteria which is an immediate corollary of monotonicity. ∎