# EXPLANATORY BELIEF ASCRIPTION

## NOTES AND PREMATURE FORMALIZATION

Kurt Konolige
Artificial Intelligence Center
Center for the Study of Language and Information
SRI International
333 Ravenswood Ave.
Menlo Park, California 94025/USA

## 1 Overview

In this paper we discuss the problem of ascribing beliefs to an agent, given partial knowledge of his beliefs. The particular kind of ascription we are interested in we call explanatory ascription, since it ascribes beliefs to an agent as a means of explaining the beliefs we already know he has.

We explore two approaches to explanatory ascription. In the first, we develop a model of belief called the derivational model, in which the derivation of one belief from another is made explicit. This model and its proof theory are formalized and used to solve a variation of the Wise Man Puzzle. For comparison, a second approach using an abductive framework and a standard modal logic of belief is developed. This approach leads to weaker conclusions than the derivational model, because closure conditions on derivations cannot be stated. On the other hand, the representational power of the two approaches differs, the abductive system being more expressive with respect to disjunctive information about belief, and the derivational model allowing nonmonotonic reasoning by the believer.

The next section of the paper gives some background in the area of belief ascription. The third section describes the problem of explanatory ascription, and some general properties that a successful approach should have. The fourth and fifth sections develop the derivational model and its proof theory, while the sixth describes an abductive framework and compares it to the derivational approach. Finally, we discuss some extensions to the language of the derivational model.

## 2 Belief ascription

Formalizations of belief are useful in AI for building systems that can represent and reason about the beliefs of other agents in cooperative situations. A typical example is an intelligent user interface to a database: an agent (the *user*) queries the system about a topic, and the system should respond in a helpful manner, making communication with the user efficient. It is well-known that extensive knowledge of the user's beliefs and intentions is required for this task [Pollack, 1986]. Since explicit communication of beliefs and intentions is limited, the system must implicitly ascribe them to the user. Here we concentrate on the process of belief ascription. Existing formalizations of belief can to some extent represent several types of ascription, which we call *deductive*, *closure*, and *analogical* ascription, but are inadequate for another type which we dub *explanatory* ascription. In this paper we develop and formalize a preliminary theory of explanatory belief ascription.

As an example of the various types of ascription, consider a simplified version of the Wise Man Puzzle with only two wise men. Each has a white hat, and it is common knowledge that there is one white hat. The first wise man says "I don't know the color of my hat," and the second wise man, on hearing this, says "My hat must be white."

As observers, we can ascribe to the second wise man the belief that his hat is white from our knowledge of his other beliefs: common knowledge of the existence of one white hat, knowledge of perceptual capabilities, and knowledge of the first wise man's ignorance of his own hat's color. This ascription is *deductive* in a suitable modal logic of belief[1]: given all of the facts above expressed in the logic, it follows that the second wise man believes his hat is white[McCarthy, 1978]. We use deduction within the belief logic to reason about the conclusions an agent must draw from an initial set of beliefs.

Deductive ascription can also be used to infer that an agent does *not* believe some fact. In the Wise Man Puzzle, the second wise man reasons that the first wise man does not believe his own hat is white, and thus can not believe that the second wise man's hat is black. More simply, if an agent believes a proposition, he does not believe its negation[2].

Another type of reasoning about ignorance is to be able to conclude that the first wise man does not know the color of his own hat, given his beliefs in the initial situation. This kind of reasoning requires a closure assumption about beliefs — the only beliefs the first wise man has that are relevant to his hat color are those presented in the puzzle. Belief ascription by closure is complicated to formalize, but recent approaches involving autoepistemic logic seem to be successful [Levesque, 1987].

A third type of ascription of belief is *analogical* in nature. The second wise man must reason that the perceptual capabilities of the first wise man are similar to his own, i.e., that he will recognize a white hat when he sees one, and form an appropriate belief. Analogical ascription is difficult to formalize, because it is only plausible: other information might cause the analogical ascription to be retracted. This type of ascription is widely used in plan recognition to make assumptions about another agent's knowledge of the effects of actions [Pollack, 1986, Konolige and Pollack, 1989].

These types of ascription all have a common core: they involve reasoning about the forward inferential connections of the beliefs of an agent. That is, they involve reasoning of the form: if an agent believes $\alpha$, and $\beta$ follows from $\alpha$, then the agent must believe $\beta$. Even ascription by closure is a variation on this theme — there is no $\alpha$ believed by the agent such that $\beta$ follows from $\alpha$, therefore the agent does not believe $\beta$. By contrast, explanatory belief ascription is *abductive* in nature: one searches for a plausible explanation for the beliefs an agent is known to have. We discuss this process in the next section.

## 3 Explanatory belief ascription

Consider a variation of the Wise Man Puzzle for two wise men, that we will call the Easy Wise Man Puzzle. Suppose that the second wise man's hat is black, and thus the first wise man says "I know that my hat is white." The second wise man then says, "My hat must be black." How can we account for this conclusion? Intuitively, the second wise man must reason about what gave

[1] The simplest such logic, $K$, will do here.

[2] This is assuming an agent's beliefs to be consistent, which requires a belief logic with the modal axiom $D$.

rise to the first wise man's belief, that is, what other beliefs it must be based on. In this case, a plausible candidate is the belief that second wise man's hat is black. We call this kind of reasoning *explanatory* ascription, because it ascribes beliefs as a means of explaining the presence of other beliefs.

What are the general characteristics of explanatory ascription? Here we discuss some of its properties, as a guide to developing a formalization in the next section.

1. The ascription of belief obviously depends on what we take to be a model of belief. Without arguing for their appropriateness, we make two related assumptions. The first is that beliefs are related by derivation, that is, on the basis of certain beliefs, an agent will derive other beliefs. The aim is to keep this assumption as abstract as possible, so no particular symbol system or computational mechanism is assumed; but the model of belief must have some way of stating derivation relations. The second assumption is that a special set of beliefs, the primitive beliefs, acts as the foundation for all derived beliefs. This view is similar to that of the Truth Maintenance System [Doyle, 1979], in which all propositions are justified on the basis of some set of primitive propositions that need no justification.[3] The set of primitive beliefs will often include those that are based on observation.

2. Explanatory ascription is a kind of abductive inference, in that it involves reasoning from an observed belief of an agent back to the way the belief arose or was derived. In this respect it differs from the other types of ascription noted above. Several consequences should be noted.[4]

First, explanations should be minimal in some sense. For belief ascription, we want to ascribe just enough primitive beliefs to account for the observed facts, and no more. In this respect explanatory ascription differs from other types of forward reasoning ascription, which are generally additive. For instance, if beliefs $\beta_1$ and $\beta_2$ are both derivable from the belief $\alpha$, and they are consistent with each other, then it is reasonable to ascribe belief in both of them. On the other hand, explanatory inference is not additive in this manner, but competitive. If $\alpha_1$ and $\alpha_2$ are both explanations for $\beta$, then even if they are mutually consistent, it would be unwise to ascribe both of them to an agent believing $\beta$.

The competitive nature of explanatory ascription is similar to that in the plan recognition process. In ascribing intentions to an agent whose actions are being observed, we seek to ascribe fragments of a plan that would connect the observed actions with the imputed goals of an agent [Konolige and Pollack, 1989]. Alternative plan fragments compete with one another as explanations of the observed actions. Local cues, such as preferences for one type of plan fragment over another, can be used to choose among the alternatives.

A similar story can be told in explanatory ascription of belief. Here the inferential connections between the beliefs, as well as beliefs themselves, are being ascribed. So, for example, in the case of the Easy Wise Man Puzzle, the second wise man ascribes to the first the belief in $b_2$, and also the inferential connection between this belief and the conclusion $w_1$.[5]

[3] The foundational theory of belief has an interesting history in the philosophy literature, and there are some compelling arguments against its full application (see, for example, [Harman, 1986]). Nevertheless it is a useful approximation.

[4] In other accounts of abduction, the properties we cite here are also recognized, e.g., [Levesque, 1989, Poole, 1988, Reiter, 1987]).

[5] $b_i$ and $w_i$ are convenient abbreviations for the propositions *wise man i's hat is black (or white).*

A second characteristic of explanatory ascription is that there may be preferences among competing explanations. It may be much more likely, for example, that Ralph knows the combination of the lock because he saw someone else open it, rather than guessing it at random. Often this kind of preference information is based on knowing that certain derivations are more likely than others. Thus it is useful to have a language in which both beliefs and the derivational relations among them are represented.

3. Distinguishing between the derivational capabilities of an agent and his "factual" beliefs is important. For example, in the Easy Wise Man Puzzle, we have information about the possible derivations that the first wise man can make from his observation of the second wise man's hat. This includes *closure* information, that is, the knowledge that the only way in which he can conclude that his own hat is white is if he sees a black hat on the second wise man. On the other hand, we do not want to assume closure over factual information, since we do not have complete knowledge here.[6]

4. Belief derivation may be *nonmonotonic*. By this we mean that an agent may come to a conclusion on the basis of some proposition he does not believe, as well as those he does. This derivation is nonmonotonic in the sense that it may be retracted when the agent acquires new information. The presence of nonmonotonic derivation poses an additional challenge for explanatory ascription.

Most of the logics of belief used in AI work, including all those derivative of Hintikka's original possible-worlds formulation [Hintikka, 1962], do not seem adequate to account for these properties, even in an abductive framework. The main problem is that they do not explicitly represent relationships among beliefs, particularly that one belief is derived from or caused by other beliefs. Rather, the derivation of one belief from another (for an ideal reasoning agent) is implicitly given by the axioms of the belief logic. For example, an ideal agent would conclude $q$ from $p$ and $p \supset q$, and this is reflected in a normal modal logic by the fact that $B(q)$ follows from $B(p)$ and $B(p \supset q)$. There is never any need to reason explicitly about the derivation of belief: one simply starts with a set of known beliefs, and uses logical entailment in the belief logic to deduce all derived beliefs. In explanatory ascription, on the other hand, being able to reason explicitly about belief derivation is necessary, both for deciding among competing explanations, and in stating closure conditions. Hence the representation of derivation or causation among beliefs becomes important.

In the sequel, we develop a derivational model of belief in which the distinction between beliefs and their derivational relations is clearly drawn. From this we construct a competence theory of belief ascription, that is, we do not consider preferences among the explanations. The ascription theory is based on the simplifying assumption of a propositional belief language and monotonic derivation.

## 4 A derivational model of belief

As we have argued, an appropriate model of belief for explanatory ascription must take into account the derivational relationships among beliefs. Together with the hypothesis of a foundational theory,

[6] This distinction means that we cannot formalize the closure conditions by simply giving a set of sentences about an agent's beliefs, and then saying "the agent has no beliefs that do not follow from these sentences." In fact, this is the approach taken in [Levesque, 1987] and [Lifschitz, 1989]. In earlier work [Konolige, 1982], this author explored a different kind of closure condition relating the background knowledge of an agent to his factual beliefs.

we construct a simple derivational model of belief. First, we fix the possible model structures by defining a *frame.*

DEFINITION 4.1 *A* frame *consists of three elements: a set of propositions* $(p_1, p_2, \cdots)$*, a subset of these propositions called the* primitive *propositions, and a mapping* $\mathcal{D}$ *from sets of propositions to propositions. An element of* $\mathcal{D}$ *is called a* derivation, *and can be written as* $p_1, \cdots, p_n \mapsto q$*. The propositions* $p_i$ *are called the antecedents, and* $q$ *the conclusion. One proposition,* $\perp$*, is distinguished as the contradictory proposition, and is never primitive.*

A frame gives the possible beliefs and derivations among beliefs that an agent can have. A model, based on a frame, describes the beliefs of an agent in a particular situation.

DEFINITION 4.2 *A* derivational model *of belief over a frame* $\langle P, Prim, \mathcal{D}\rangle$ *is a tuple* $\langle B, D\rangle$*. The* belief set $B$ *is a subset of* $P$*, and the* derivation set $D$ *is a subset of* $\mathcal{D}$*. The following conditions must hold:*

1. *The contradictory proposition* $\perp$ *is not in* $B$*.*
2. *For every derivation* $p_1 \cdots p_n \mapsto q$ *of* $D$*, all propositions* $p_i$ *and* $q$ *are in* $B$*.*
3. *If* $d = p_1 \cdots p_n \mapsto q$ *is a derivation of* $\mathcal{D}$*, and all* $p_i$ *are in* $B$*, then* $d \in D$*.*
4. *Every element of* $B$ *is the root of a tree over* $D$ *whose leaves are in Prim.*

Informally, a model gives the beliefs of an agent (component $B$), together with the derivational structure of those beliefs (component $D$). The conditions on $B$ and $D$ ensure that the set of beliefs is closed under well-founded derivation. The exclusion of $\perp$ means that the beliefs are noncontradictory. The definition of the derivational model is similar to admissible labelings of a TMS [Reinfrank *et al.*, 1989] with only monotonic rules.

It is straightforward to give models for the beliefs of the first Wise Man in the Easy Wise Man Puzzle. The propositions are $w_1$, $b_1$, $w_2$, and $b_2$. For the first Wise Man, $w_2$ and $b_2$ are primitive because they are observable. Since he knows that at least one hat is white, his derivational relation $\mathcal{D}$ is:

$$\begin{aligned} b_1 &\mapsto w_2 \\ b_2 &\mapsto w_1 \\ b_1, w_1 &\mapsto \perp \\ b_2, w_2 &\mapsto \perp \end{aligned} \qquad (1)$$

There are three possible models using this frame:

| *model* | $B$ | $D$ |
|---|---|---|
| $m_1$ | $\emptyset$ | $\emptyset$ |
| $m_2$ | $b_1, w_2$ | $b_1 \mapsto w_2$ |
| $m_3$ | $b_2, w_1$ | $b_2 \mapsto w_1$ |

The model in which both $b_1$ and $b_2$ are in the belief set leads to a contradiction.

The derivational model can be considered as a further development of the Deduction Model of belief [Konolige, 1986]. In contrast to possible-world model, the Deduction Model considers belief to be sentence-like data structures in the cognitive structure of an agent, who has a deductive apparatus for infering one belief from another. The most significant departure here is the inclusion of derivation as an explicit structural element of the model. Like the deduction model, the derivational model is free from the problem of logical omniscience, in which it is assumed that an agent knows all the logical consequences of his beliefs. However, an agents' beliefs are assumed to be closed under derivation — but the derivation mapping $\mathcal{D}$ of a frame may not be logically complete.

Another way in which the derivational model differs significantly is the inclusion of the derivation mapping of a frame. This mapping specifies all of the possible ways in which a belief can be derived from its fellows, and as such is a type of closure. The mapping may have an infinite number of members, but it can also be finite, or the number of derivations with a given proposition as the conclusion may be finite. With the appropriate language for describing models, it is very easy to express the closure conditions that prevail in many belief representation situations.

## 5 A language and proof system

Our formalization of the derivational model of belief will proceed in several parts. First we present a language for talking about belief, and give its semantics relative to the model. Then we define the notion of logical implication relative to a set of premises and a frame. Finally, we construct an axiomatic system and prove it sound and complete with respect to logical implication, when the frame is finite.

### 5.1 A belief language

A propositional language for belief, **B**, is defined relative to a derivational frame $\langle P, Prim, \mathcal{D}\rangle$. It contains:

- A set of modal atoms B$p$, where $p \in P$. $p_1, \cdots, p_n \mapsto q \in \mathcal{D}$.
- A set of *ordinary* propositional atoms, and the boolean connectives.

The semantics is straightforward. An interpretation consists of a truth assignment $v$ for all of the ordinary atoms, along with a derivational structure $m = \langle B, D\rangle$ over a frame $\mathcal{F}$. The normal rules for boolean connectives hold; the atoms are interpreted as follows:

1. $m, v \models_{\mathcal{F}} \phi$, for $\phi$ ordinary, iff $v(\phi) =$ true.
2. $m, v \models_{\mathcal{F}}$ B$p$ iff $p \in B$.

Note that interpretations are always defined relative to a frame. In the standard way, a set of sentences $\Gamma$ of **B** defines the collection of models for which all of $\Gamma$ are true, and a sentence $\phi$ true in all these models is a logical consequence of $\Gamma$; we write $\Gamma \models_{\mathcal{F}} \phi$.

### 5.2 Inference

Our knowledge about an agent is given by two collections of belief propositions and derivation relations. One collection is our knowledge of the possible belief derivations and primitive propositions an agent might have: the frame. For example, the frame for the Easy Wise Man Puzzle, from Equation 1 and the paragraph preceding it, gives the possible derivations of the first Wise Man from the primitive propositions $b_2$ and $w_2$. As noted above, an important part of the information given is the implicit closure condition: there are no other derivations the agent could make.

The other collection is our initial knowledge of the agent's beliefs, a finite set of sentences of B called the *base* or *premise set*. The base set $\{\mathrm{B}w_1, \neg\mathrm{B}b_1\}$ describes the first Wise Man as believing (at least) $w_1$ and not believing $b_1$. A base set may be incomplete, in that it may not contain all of the beliefs that are logical consequences of its members relative to the derivational semantics (the base set just given is incomplete with respect to $b_2$ and $w_2$).

Given a frame $\mathcal{F}$ and a base set $O$, what conclusions should we come to? At the least, we should conclude everything that is true in all models of $\mathcal{F}$ and $O$. That is, we conclude:

$$\{\phi \mid O \models_{\mathcal{F}} \phi\}. \tag{2}$$

This is the set of logical consequences of $O$, given a fixed frame $\mathcal{F}$.

As an example, consider the frame with primitive propositions $\{a_1, a_2\}$ and derivations

$$a_1 \mapsto p,\ a_2 \mapsto p,\ a_2 \mapsto q.$$

Suppose the base set is $\{\mathrm{B}a_1\}$. There are three models:

| *model* | *B* | *D* |
|---|---|---|
| $m_1$ | $a_1, p$ | $a_1 \mapsto p$ |
| $m_2$ | $a_2, p, q$ | $a_2 \mapsto p, a_2 \mapsto q$ |
| $m_3$ | $a_1, a_2, p, q$ | $a_1 \mapsto p, a_2 \mapsto p, a_2 \mapsto q$ |

From these we can conclude $\mathrm{B}a_1 \vee \mathrm{B}a_2$, $\neg\mathrm{B}a_1 \supset \mathrm{B}q$, $\neg\mathrm{B}q \supset \neg\mathrm{B}a_2$, $\neg\mathrm{B}a_2 \supset \neg\mathrm{B}q$, etc. Note that, because of the implicit closure condition of the frame, we have greatly increased the ability to infer ignorance on the part of the agent.

As another example, consider the same frame with the base set $O = \{\mathrm{B}a_1, \mathrm{B}q\}$. There are two models, $m_2$ and $m_3$ of the previous example. We should conclude $\mathrm{B}a_2$, but not $\neg\mathrm{B}a_1$, since $a_2$ is a belief in $m_3$. Note that this is different from the conclusions we get by taking the models of $O$ that are minimal in the primitive propositions believed: this would be $m_2$ alone, and $\neg\mathrm{B}a_1$ would be a conclusion.

Finally, consider the Easy Wise Man example of Equation 1. For the base set $\{\mathrm{B}w_1\}$, there is only one model, namely that for which $\mathrm{B}b_2$ holds.

Although it may not be obvious, the derivational semantics are sufficient for concluding the minimal abductive consequences of a base set. We prove this in the next section, where the derivational semantics is compared to the abductive framework.

## 5.3 Proof theory

The proof theory is a propositional system, with additional axioms for the modal atoms. Define the set $\mathrm{Ax}(\langle P, \mathit{Prim}, \mathcal{D}\rangle)$ as:

1. $\neg\mathrm{B}\perp$.

2. For every $p_1, \cdots, p_n \mapsto q \in \mathcal{D}$, $\mathrm{B}p_1 \wedge \cdots \wedge \mathrm{B}p_n \supset \mathrm{B}q \in \mathrm{Ax}(\langle P, Prim, \mathcal{D}\rangle)$.
3. Let $q \in P$, $q \notin Prim$, and $q \neq \perp$. If the number $N$ of derivations of $q$ in $\mathcal{F}$ is finite, then $\mathrm{B}q \supset \bigvee_{i=1}^{N}(\mathrm{B}p_1^i \wedge \cdots \wedge \mathrm{B}p_{k_i}^i) \in \mathrm{Ax}(\langle P, Prim, \mathcal{D}\rangle)$.

The first axiom states that beliefs are not contradictory. The second item contains axioms that state how beliefs are inferred by an agent via derivations. The third is basically Clark's completion of the first set of axioms, stopping at the primitive propositions. If a particular proposition has an infinite number of derivations, then we can't form the completion in the language **B**.

Define the theorems of a base set $O$ on a frame $\mathcal{F}$ by:

$$O \vdash_{\mathcal{F}} \phi \quad \text{if and only if} \quad O \cup \mathrm{Ax}(\mathcal{F}) \vdash \phi. \tag{3}$$

We can prove the following soundness theorem:

THEOREM 5.1 *$O \models_{\mathcal{F}} \phi$ if $O \vdash_{\mathcal{F}} \phi$.*

There are conditions under which the converse is also true, so that the axiomatic system is complete.

THEOREM 5.2 *Suppose $\mathcal{F}$ is an acyclic frame such that every proposition is the conclusion of only a finite number of derivations. Then $O \vdash_{\mathcal{F}} \phi$ if $O \models_{\mathcal{F}} \phi$.*

For the Easy Wise Man frame, we get the following set of axioms:

$$\mathrm{Ax}(\mathcal{F}) = \{ \begin{array}{l} \mathrm{B}b_1 \equiv \mathrm{B}w_2 \\ \mathrm{B}b_2 \equiv \mathrm{B}w_1 \\ \neg(\mathrm{B}b_1 \wedge \mathrm{B}w_1) \\ \neg(\mathrm{B}b_2 \wedge \mathrm{B}w_2) \quad \} \end{array} \tag{4}$$

Besides knowledge of the derivational frame, the second wise man knows the following about the first wise man: he believes his own hat is white ($\mathrm{B}w_1$), and he can faithfully observe the color of the second wise man's hat ($b_2 \supset \mathrm{B}b_2$, $w_2 \supset \mathrm{B}w_2$). From these and the frame axioms it follows that:

$$\mathrm{Ax}(\mathcal{F}) \cup \{Bw_1, b_2 \supset \mathrm{B}b_2, w_2 \supset \mathrm{B}w_2\} \vdash b_2 \,.$$

This is the solution to the Easy Wise Man Puzzle using the model of derivational belief.

## 6 An abductive framework

In this section we compare the derivational model with a standard modal belief logic in an abductive framework. We show that, under suitable assumptions, the derivational model produces all the conclusions of the standard logic, but that the converse is not true.

We define a standard modal language $\mathbf{B}'$ based on the propositions $P$ of a frame. The sentences of $\mathbf{B}'$ are all boolean combinations of $P$ and the modal atoms $\mathrm{B}\phi$, where $\phi$ is an ordinary (nonmodal) sentence of $\mathbf{B}'$. We are not interested in the complications of nested modal operators here. We have $\mathbf{B} \subset \mathbf{B}'$, since $\mathbf{B}$ excludes those modal atoms whose arguments are boolean combinations of the propositions $P$.

The set $Pr$ is defined as $\{\mathrm{B}\phi \mid \phi \in \mathit{Prim}\}$, that is, the set of primitive belief atoms. Take the logic to be propositional $KD$, that is, the simplest belief logic together with the axiom $D = \neg\mathrm{B}\perp$ stating that beliefs are consistent. In the abductive framework, there is background information $\Sigma \subset \mathrm{B}'$ about the world and the agent's beliefs. The background information might contain statements about the connections among the agent's beliefs, or between the agent's beliefs and the world. There is also a set of observations $O \subset \mathrm{B}$ about the agent's beliefs, from which further beliefs are to be ascribed.

An *explanation* of the observations is a set $A \subset Pr$ such that

1. $\Sigma \cup A \models_{KD} O$.

2. $A$ is consistent in $KD$.

3. $A$ is minimal.

A *cautious explanation* is the disjunction of all the explanations, that is, $\bigvee_i A_i$. This is the minimum we can conclude in the abductive framework.

As an example consider the Easy Wise Man Puzzle. The background information, from the second wise man's point of view, consists of:

$$\begin{array}{ll}
\mathrm{B}(w_1 \vee w_2) & \text{the first wise man believes at least one hat is white} \\
w_1 \vee w_2 & \text{and so it is} \\
w_2 \supset \mathrm{B}(w_2) & \text{he can observe the second wise man's hat} \\
b_2 \supset \mathrm{B}(b_2) & \\
\mathrm{B}(b_1 \equiv \neg w_1) & \text{he believes black and white are mutually exclusive} \\
\mathrm{B}(b_2 \equiv \neg w_2) & \\
b_1 \equiv \neg w_1 & \text{as indeed they are} \\
b_2 \equiv \neg w_2 &
\end{array} \tag{5}$$

From the observed fact $\mathrm{B}w_1$, there is only one possible explanation (which is also the cautious explanation), $\mathrm{B}b_2$. From this and the background information it follows that $\mathrm{B}w_1$, and also $\neg\mathrm{B}w_2$ (via the $D$ axiom) and hence $b_2$. So by using a standard belief logic in an abductive framework, the second wise man can infer the color of his own hat.

The abductive framework actually produces only a subset of the conclusions of the derivational model: conclusions about the non-beliefs of the agent are lacking. Consider the same example, only with the observation set $O = \{\neg\mathrm{B}b_2\}$. The cautious explanation is the tautology $\neg\perp$, so we can conclude only what follows in $KD$ from $\Sigma$ and $O$; in particular, we cannot conclude $\neg\mathrm{B}w_1$, which is a consequence of the derivational model. The difference lies in the ability to state closure of the possible derivations in the derivational model. In this model, the only way in which $w_1$ could be derived by the agent is via the belief in $b_2$, and so not believing $b_2$ means he doesn't believe $w_1$.

It is possible to prove that, under the condition of complete derivation, the derivational model produces all of the conclusions of the standard model in the abductive framework. Define a complete set of derivations relative to a background theory as:

DEFINITION 6.1 $\mathrm{Comp}(\Gamma, P, Q)$ *is the set of all derivations* $p_1, \cdots, p_n \mapsto q$ *such that*

*1.* $p_i \in P$ *and* $q \in Q$*,*

2. *$p_1, \cdots p_n$ is a minimal subset such that $\Gamma \models_{KD} \mathrm{B}p_1 \wedge \cdots \wedge \mathrm{B}p_n \supset \mathrm{B}q$,*
3. *either $\mathrm{B}p_1 \wedge \cdots \wedge \mathrm{B}p_n$ is consistent with $\Gamma$, or $q = \perp$.*

Divide the background theory $\Sigma$ into two parts: a set $\Sigma_d$ whose sentences contain modal atoms not in B (that is, with complex propositional arguments), and the rest $\Sigma_b$. We use $\Sigma_d$ to generate derivations.

THEOREM 6.1 *Let $\Sigma = \Sigma_d \cup \Sigma_b$ be as defined above, and let $A$ be the cautious explanation of some observation set $O$ relative to $\Sigma$. Suppose $\Sigma_d$ consists only of modal atoms and their negations. Let $\mathcal{F}$ be the frame $\langle P, Prim, \mathrm{Comp}(\Sigma_d, Prim, P)\rangle$. Then*

$$\Sigma_b \cup O \models_{\mathcal{F}} A\ .$$

## 7 Extensions

We briefly discuss several extensions to the formalization. These are preliminary ideas, and need further exploration.

### 7.1 Language

The belief language B is impoverished with respect to the model, since we assume a fixed frame, rather than allowing statements about the frame in the language. In part this is to distinguish knowledge of belief derivations from knowledge of factual belief, and in part to make it easy to state closure conditions on derivations. A reasonable extension would be to eliminate the assumption of a fixed frame, and expand the language to include statements about derivations, perhaps of the form $\mathrm{Der}(p_1, \cdots, p_n; q)$ to indicate that $p_1, \cdots, p_n \mapsto q$ is a derivation. To achieve closure over the derivations given by a set of premises, we could make inferences with respect to derivational models of the premises with *minimal frames*. This is equivalent to the present approach when the premises only contain the modal atoms Der as sentences.

### 7.2 Nonmonotonic belief derivation

A slight change in the definition of derivations allows us to add nonmonotonic derivations to the model. Instead of $p_1, \cdots p_n \mapsto q$, we take derivations to be $p_1 \cdots p_n; r_1 \cdots r_l \mapsto q$, where the $r_i$ are the nonmonotonic antecedents of the derivation. The conditions on the model (Definition 4.2) must be modified:

2. For every derivation $p_1 \cdots p_n; r_1 \cdots r_l \mapsto q$ of $D$, all propositions $p_i$ and $q$ are in $B$, and $r_j$ are not in $B$.

3. If $d = p_1 \cdots p_n; r_1 \cdots r_l \mapsto q$ is a derivation of $\mathcal{D}$, and all $p_i$ are in $B$, and all $r_j$ are not in $B$, then $d \in D$.

We make similar modifications in the proof theory, adding negative belief literals in the antecedent of the implications in $\mathrm{Ax}(\mathcal{F})$. Nothing else need be changed: the soundness and completeness theorem go through as before.

Surprisingly, we are able to formalize quite simply the behavior of an agent that does nonmonotonic reasoning. However the belief language is still restrictive, limiting us to the case of a TMS-reasoner, a subset of a full default-logic type of reasoner [Reinfrank *et al.*, 1989].

## 8 Conclusion

The theory of belief ascription presented here accounts for some of the characteristic interaction of explanatory and deductive ascription. Previous formal work on belief ascription in AI has not addressed the problem of explanatory ascription.

We have developed an elaboration of the Deduction Model of belief, by making explicit the nature of derivation among beliefs. The advantages of this model are that it allows us to state closure conditions on the derivations in a straightforward manner, and keep them separate from closure conditions on "factual" beliefs, which are usually not desired. We have developed a proof theory for the model, and shown it to be sound, and complete under certain restrictions on finiteness in the derivational structure.

In comparison with a standard modal logic of belief in an abductive framework, the derivational model has good and bad points. On the one hand, for a restricted background theory, it gives all of the results of the abductive system, and in addition allows conclusions based on the closure of derivations that are not available in the abductive system. On the other hand, the expressivity of the language B is somewhat impoverished. Perhaps a good solution here is to expand the expressivity of B to talk about derivations; but then the problem of defining closure of these derivations appears.

Finally, we have indicated how the derivational model can deal with belief ascription when the agent is a nonmonotonic reasoner, something that cannot be done in the standard abductive framework.

## 9 Acknowledgements

I would like to thank David Israel and Martha Pollack for discussing these issues with me.

## References

[Doyle, 1979] John Doyle. A truth maintenance system. *Artificial Intelligence*, 12(3), 1979.

[Harman, 1986] Gilbert Harman. *Change in View.* The MIT Press, Cambridge, Massachusetts, 1986.

[Hintikka, 1962] Jaako Hintikka. *Knowledge and Belief.* Cornell University Press, Ithaca, New York, 1962.

[Konolige and Pollack, 1989] Kurt Konolige and Martha Pollack. Ascribing plans to agents: Preliminary report. In *submitted to AAAI*, Detroit, Michigan, 1989.

[Konolige, 1982] Kurt Konolige. Circumscriptive ignorance. In *Proceedings of the American Association of Artificial Intelligence*, Pittsburgh, Pennsylvania, 1982. Carnegie-Mellon University.

[Konolige, 1986] Kurt Konolige. *A Deduction Model of Belief.* Pitman Research Notes in Artificial Intelligence, 1986.

[Levesque, 1987] Hector J. Levesque. All I know: an abridged report. In *Proceedings of the American Association of Artificial Intelligence.* Seattle, Washington, 1987.

[Levesque, 1989] Hector J. Levesque. A knowledge-level account of abduction. In *Proceedings of the International Joint Conference on Artificial Intelligence.* Detroit, Michigan, 1989.

[Lifschitz, 1989] Vladimir Lifschitz. Between circumscription and autoepistemic logic. In *Proceedings of the First International Conference on Knowledge Representation and Reasoning*, Toronto, Ontario, 1989.

[McCarthy, 1978] John McCarthy. Formalization of two puzzles involving knowledge. Unpublished note, 1978.

[Pollack, 1986] M. E. Pollack. *Inferring Domain Plans in Question-Answering.* PhD thesis, University of Pennsylvania, 1986.

[Poole, 1988] David Poole. A methodology for using a default and abductive reasoning system. Technical report, Department of Computer Science, University of Waterloo, Waterloo, Ontario, 1988.

[Reinfrank *et al.*, 1989] M. Reinfrank, O. Dressler, and G. Brewka. On the relation between truth maintenance and autoepistemic logic. In *Proceedings of the International Joint Conference on Artificial Intelligence.* Detroit, Michigan, 1989.

[Reiter, 1987] Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32, 1987.