Mental State Ascription Using Dynamic Logic

Michal Sindlar and Mehdi Dastani and John-Jules Meyer

Abstract. In situations where the behavior of a system must be interpreted because its state is not accessible, it is useful to explain observed behavior in mentalistic terms. This paper presents a formalism based on propositional dynamic logic to model ascription of beliefs, goals, or plans on grounds of observed actions. The formalism is used to provide semantics for an existing approach to abducing the mental state of an observed agent; in doing so it is shown how behavior-producing rules can be given different explanatory interpretations.

1 Introduction

The phenomenon of ascribing mental states to an entity in order to account for its observed behavior, to which Dennett’s theory of the Intentional Stance [7] gives a philosophical underpinning, has been studied extensively in the social sciences [13]. In A.I. plan/activity/intention recognition and user modeling are well-known domains that approach (aspects of) this phenomenon formally, which has wide applicability [3]. The game domain offers an ideal testbed for research in this field [12], because different sorts of games call for virtual characters that interact with each other’s rule libraries, such that they can (defeasibly) infer others’ mental states and act on that information. The former occurs when virtual characters interact with heterogeneous others, the latter when virtual characters are given partial information about other characters’ mental state in order to maintain believability.

In earlier work we presented an approach to abducing the mental states of BDI-based software agents on grounds of their observed actions [17]. Development of believable virtual characters is a possible application, and we envisage characters is a possible application, and we envisage characters being programmed as BDI-based software agents to have access to a set of ascribed mental states under different perceptory conditions that represent partial observability. The approach focuses on syntactic matching of observed and observable sequences, but lacks semantics for actions and ascription.

In the current paper we present a formalism based on PDL (propositional dynamic logic [9]) with poor tests, that allows for characterizing the procedures presented in [17], where observed behavior is that of a 2APL-like agent which follows a non-interleaving execution strategy. The formalism also allows for specifying ascribed mental states in relation to observed behavior in general, such that the ‘theory of mind’ of an observer can be modeled (see also [18]).

The paper is structured as follows. Section 2 briefly reiterates mental state abduction [17]. Section 3 defines the formalism which is used in Section 4 to characterize mental state abduction, and to extend upon this characterization in various ways by providing different explanatory interpretations. Sections 5 and 6 conclude, respectively, with a discussion of related work and reflection upon this and future research.

2 Mental State Abduction

The behavior of agents is taken to be produced by goal-directed rules of the form $\gamma \prec \beta \pi$, which state that to achieve goal $\gamma$, if $\beta$ is believed, the plan $\pi$ is appropriate and can be selected. The BNF grammar of rules is given below.

$\langle library \rangle \::= \langle pgrule \rangle +$

$\langle pgrule \rangle \::= \langle bq \rangle " \cdot \cdot " \langle bq \rangle " \cdot \cdot " \langle plan \rangle$

$\langle plan \rangle \::= \langle action \rangle | \langle test \rangle | \langle seqplan \rangle | \langle tifplan \rangle | \langle whplan \rangle$

$\langle action \rangle \::= \langle ifplan \rangle$" test"$\langle plan \rangle$

$\langle test \rangle \::= \langle value \rangle | \langle predicate \rangle$" then"$\langle plan \rangle$" else"$\langle plan \rangle$

$\langle seqplan \rangle \::= \langle plan \rangle$";"$\langle plan \rangle$

$\langle tifplan \rangle \::= \langle if \rangle$" test"$\langle plan \rangle$

$\langle whplan \rangle \::= \langle while \rangle$" test"$\langle do \rangle$"$\langle plan \rangle$

Formal semantics of rule interpretation is omitted, but see [6]. Some assumptions are made, which are justified in the context of, e.g., virtual character design: observed behavior of agents stems from a known library of rules; only primitive actions are observable; agents have a non-interleaving execution strategy [1]; and, plans can be dropped before completion.

It is assumed here that elements of the above rule definition can be translated into logical expressions (see [1] for procedures to do so). Let $p \in \text{Atom}$ be an atomic proposition, $\sim p$ the negative literal of $p$, and $\alpha \in \text{Act}$ a primitive action. The language $\mathcal{L}_0$, represented below by $\psi$, is then a basic propositional language with positive and negative literals, composed by conjunction or disjunction.

$\psi \ ::= p | \sim p | \top | \bot | \psi_1 \land \psi_2 | \psi_1 \lor \psi_2$

$\pi \ ::= \alpha | \mathbf{B}\psi | \mathbf{G}\psi | \pi_1; \pi_2 | \pi_1 + \pi_2 | \pi^*$

The operator $\mathbf{B}$ represents a test action on propositions of $\mathcal{L}_0$, $\mathbf{G}$ represents sequential composition, $\cdot$ non-deterministic choice, and $^*$ iteration. In the following it is assumed that for rules of the form $\gamma \prec \beta \pi$ it holds that $\gamma, \beta \in \mathcal{L}_0$ and $\pi \in \mathcal{L}_n$. 
2.1 Explanation of Observed Behavior

To explain observed behavior in mentalistic terms the abductive syllogism can be followed, which states that given the rule $\phi \rightarrow \psi$ and an observed fact $\psi$, the abducible fact $\phi$ can (defeasibly) be inferred. It is slightly adapted for the particular case of inferring beliefs and goals in relation to plans and observed behavior, as shown below.

\[
\begin{align*}
&\text{from } \text{goal}(\gamma) & \& \text{belief}(\beta) \implies \text{select}(\pi) \\
&\text{and } \text{observed}(\alpha_1 \ldots \alpha_n) & \& \text{related}(\alpha_1 \ldots \alpha_n, \pi) \\
&\text{infer } \text{goal}(\gamma) & \& \text{belief}(\beta)
\end{align*}
\]

This syllogism states that if an implicatory relation between a mental state (goal/belief pair) and selection of some plan is supposed, and actions are observed which can be somehow related to this plan, then the mental state can be abduced.

To relate observed primitive actions to a plan, observable sequences which are generated by this plan can be considered [17]. Given $\alpha \in \text{Act}$ let $\delta := \alpha_1; \delta_2$ be the definition of the percept language $L_\alpha \subseteq L_\alpha$ with typical element $\delta$. The function $OS : L_\alpha \rightarrow \varphi(L_\Delta)$ translates complex expressions, which include tests and looping or branching constructs, to a set of sequences of observable actions.

\[
\begin{align*}
&OS(\alpha) = \{ \alpha \} \\
&OS(\delta?) = \emptyset \\
&OS(\pi_1; \pi_2) = OS(\pi_1) \cup OS(\pi_2) \\
&OS(\pi_1 + \pi_2) = OS(\pi_1) \cup OS(\pi_2) \\
&OS(\pi^*) = \bigcup_{n \in \mathbb{N}_1} OS(\pi^n), \text{ where } \pi^1 \triangleq \pi \land \pi^{n+1} \triangleq \pi; \pi^n
\end{align*}
\]

The composition operator $\bullet : \varphi(L_\Delta) \times \varphi(L_\Delta) \rightarrow \varphi(L_\Delta)$ takes arguments $\Delta_1, \Delta_2 \subseteq L_\Delta$ and maps them to $\{ \delta_1; \delta_2 \mid \delta_1 \in \Delta_1, \delta_2 \in \Delta_2 \}$ if $\Delta_1 \neq \emptyset \land \Delta_2 \neq \emptyset$, to $\Delta_1 \land \Delta_2 = \emptyset$, or to $\emptyset$ otherwise. Note that $\pi^*$ is not considered for $\pi^*$, because our interest is only in observables.

If every action of an agent is observed, then a possible relation to serve in the abductive syllogism is to check whether an observed action sequence matches some observable sequence of a plan, which justifies the abduction of any belief/goal pair accompanying that plan in a role as the agent’s possible mental state. But if abduction is to be an online process, or if plans are not necessarily finished, a more useful relation is the prefix relation $\leq$ on $L_\Delta$, defined as $\leq = \{ (\delta, \delta') \mid \delta, \delta' \in L_\Delta \}$. The abductive explanatory function $\text{abd}$, defined below, takes an action sequence and a set of arguments and maps these to goal/belief pairs if the observed sequence is a prefix of an observably trace of the corresponding plan.

\[
\text{abd}(\delta, R) = \{ (\gamma, \beta) \mid \exists \gamma - \beta | \pi \in R \exists \delta' \in OS(\pi) : \delta \leq \delta' \}
\]

Observe that the function $\text{abd}$ has the property $\forall \delta, \delta' \in L_\Delta : (\delta \leq \delta' \implies \text{abd}(\delta', R) \subseteq \text{abd}(\delta, R))$. For more details and a treatment of incomplete observation see [17].

3 Logic

This section presents the syntax, semantics, and properties of the formalism used in the next section to give model-theoretic semantics to the approach restated in the previous section.

3.1 Syntax

In the following definition of $L$, represented by $\phi$, let (given $L_0, L_1$ as defined in Section 2) $\psi \in L_0$ and $\pi \in L_1$.\n
\[
\phi := p \mid B \psi \mid G \psi \mid O \alpha \mid \neg \phi \land \phi_1 \land \phi_2 \mid [\pi] \phi \mid [\pi^*] \phi
\]

The language thus defined is a variant of PDL with converse actions and poor tests. Note that the negation sign $\neg$ in $L_0$ negates only single atoms, whereas $\land$ ranges over propositions such that from $\land$ and $\land$ other logical connectives are derived.

The language $L$ serves to describe an observer which observes and explains the actions of some agent. This perspective is important, because the operators $B$, $G$ and $O$ qualify propositions ascribed to the agent by the observer as its beliefs and goals, respectively. The operator $O$ qualifies an action of the agent as having been observed. Note that $B$, $G$

3.2 Semantics

The language $L$ is interpreted in Kripke models $\mathfrak{M}$, such that $\mathfrak{M} = \{ W, \{ R_\alpha \mid \alpha \in \text{Act} \}, \theta_B, \theta_G, \theta_I, \theta_O \}$ where $W$ is a set of states, $R_\alpha \subseteq W \times W$ the accessibility relation for primitive action $\alpha$, and the $\theta_i$’s are valuation functions. Let $\text{Lit} = \{ p, \lnot p \mid p \in \text{Atom} \}$. The valuation functions $\vartheta_B : \text{Lit} \rightarrow \varphi(W)$ and $\vartheta_O : \text{Lit} \rightarrow \varphi(W)$ assign states to literals, $\vartheta_I : \text{Atom} \rightarrow \varphi(W)$ assigns states to atoms, and $\vartheta_O : \text{Act} \rightarrow \varphi(W)$ reifies the observation of actions. The function $\vartheta : L \rightarrow \varphi(W)$ is the valuation function for propositions in $L$ which is defined on top of $\vartheta_B, \vartheta_G, \vartheta_I, \vartheta_O$ as follows, where for brevity $Q$ denotes either operator $B$ or $G$, and $\vartheta_Q$ the corresponding function $\vartheta_B$ and $\vartheta_G$, respectively, for which holds $\vartheta_Q(T) = W$ and $\vartheta_Q(\bot) = \emptyset$.

\[
\begin{align*}
\vartheta(p) &= \vartheta_I(p), \quad \vartheta(\lnot \phi) = W - \vartheta(\phi) \\
\vartheta(\phi_1 \land \phi_2) &= \vartheta(\phi_1) \cap \vartheta(\phi_2), \quad \vartheta(\Box \phi) = \vartheta_Q(\phi) \\
\vartheta_Q(\psi_1 \lor \psi_2) &= \vartheta_Q(\psi_1) \cup \vartheta_Q(\psi_2), \quad \vartheta(\Box \alpha) = \vartheta_O(\alpha) \\
\vartheta_Q(\psi_1 \land \psi_2) &= \vartheta_Q(\psi_1) \cap \vartheta_Q(\psi_2)
\end{align*}
\]

\[
\begin{align*}
\vartheta([\pi]\phi) &= \{ w \mid w' \in W : (w, w') \in \varrho(\pi) \Rightarrow w' \in \vartheta(\phi) \\
\vartheta([\pi^*]\phi) &= \{ w \mid w' \in W : (w, w') \in \varrho(\pi^*) \Rightarrow w' \in \vartheta(\phi) \}
\end{align*}
\]

The operator $^{-1}$ denotes the inverse of a relation, such that for any relation $R$, it is the case that $R^{-1} = \{ (y, x) \mid (x, y) \in R \}$. The interpretation of plans $\pi \in L_\Pi$ is based on a function $\varrho : L_\Pi \rightarrow \varphi(W \times W)$, defined as follows.

\[
\begin{align*}
\varrho(\alpha) &= R_\alpha \\
\varrho(\Box \phi(\pi)) &= \{ (w, w) \mid w \in \vartheta_Q(\Box \phi(\pi)) \\
\varrho(\pi_1; \pi_2) &= \varrho(\pi_1) \circ \varrho(\pi_2) \\
\varrho(\pi_1 + \pi_2) &= \varrho(\pi_1) \cup \varrho(\pi_2) \\
\varrho(\pi^*) &= \bigcup_{n \in \mathbb{N}_0} \varrho(\pi^n), \text{ where } \pi^0 \triangleq (\text{skip})
\end{align*}
\]

In the definition of $\varrho$, the operator $\circ$ denotes relational composition, such that for any two relations $R$ and $R'$, their composition $R \circ R' = \{ (x, z) \mid (x, y) \in R, (y, z) \in R' \}$. $\cup$ denotes set
union, and π* iteration of π, whose semantics is the reflexive transitive closure of φ(π). Zero-length iteration of π is defined as the skip action, which in this framework corresponds to \( Q(\pi) \) for \( Q \in \{ B, G \} \). Given a Kripke model \( M \), the fact that a state \( w \) satisfies a formula \( \phi \) is defined as

\[
M, w \models \phi \iff w \in \phi(\phi)
\]

\( M \models \phi \) denotes that any state in \( M \) satisfies \( \phi \), and \( \models \phi \) as a shorthand for \( M, w \models \phi \) that an action which is considered ‘observed’ is also considered as done, as opposed to multiple actions having been done simultaneously. This assumption is reflected in Property 1.

\[
(M, w \models \text{Done}(\alpha) \land \text{Done}(\alpha')) \implies (\alpha = \alpha') \quad (1)
\]

Observation of actions is formalized through the valuation function \( \mathrm{Ob} \), which assigns actions to states. It is here assumed that an action which is considered ‘observed’ is also considered done, leading to Property 2.

\[
\models \mathrm{Ob} \alpha \rightarrow \text{Done}(\alpha) \quad (2)
\]

In order to be able to talk about sequences of observed actions in this our logical language, the expression \( \text{Ob}(\delta) \) for any \( \delta \in L_{\Omega} \) is defined in terms of primitives.

\[
\text{Obs}(\alpha_1; \ldots; \alpha_n) \equiv \text{Ob} \alpha_1 \land <\alpha_1; \ldots; \alpha_n> \land
\]

\[
[\alpha_n^{-1}] \text{Ob} \alpha_{n-1} \land [\alpha_{n-1}]^{-1} \text{Ob} \alpha_{n-2} \land \ldots \land [\alpha_2; \ldots; \alpha_{n-1}]^{-1} \text{Ob} \alpha_1
\]

Furthermore, observed agents are taken to be rational, such that they are taken to not already believe the facts which they still have as goal to achieve, as stated in Property 3.

\[
\models \text{G} \psi \rightarrow \neg \text{B} \psi
\]  

(3)

It should be noted that ascription by means of the operators \( B \) and \( G \), as presented in Section 3.2, is paracomplete [14]. This means that for any \( p \in \text{Atom} \) it can be the case that neither \( p \) nor its literal negation \( \neg p \) hold, either of the two holds, or both hold simultaneously. Let the following expressions denote the fact that the status of a literal as ascribed belief is conclusive, inconclusive, or unknown.

\[
\begin{align*}
& B \text{Con}(p) \equiv Bp \lor \neg B\neg p, \quad B \text{Con}(\sim p) \equiv Bp \land \neg Bp \\
& B \text{Inc}(p) \equiv Bp \land B \sim p, \quad \text{likewise for } B \text{Inc}(\sim p) \\
& B \text{Unk}(p) \equiv \neg(Bp \lor B \sim p), \quad \text{likewise for } B \text{Unk}(\sim p)
\end{align*}
\]

Replacing \( B \) with \( G \) in the above definitions does likewise for goal-ascription of literals. Also, the definitions can be extended to conjuncted or disjuncted literals straightforwardly as follows: \( Q(\psi \land \psi') \equiv Q(\psi) \land Q(\psi') \) and \( Q(\psi \lor \psi') \equiv Q(\psi) \lor Q(\psi') \), where \( Q \in \{ B \text{Con}, B \text{Inc}, B \text{Unk}, G \text{Con}, G \text{Inc}, G \text{Unk} \} \).

## 4 Mental State Ascription

This section presents different interpretations of the approach of Section 2 [17]) using the formalism presented in Section 3.

### 4.1 Ascription of Abducibles

Behavioral rules of the form \( \gamma \land \beta \land \pi \) can be informally described as “the plan \( \pi \) is appropriate for achieving the goal \( \gamma \) if \( \beta \) is believed”. If an agent’s observed actions are abductively explained with respect to some set of rules, then a model should reflect the relation stated by those rules to hold between behavior and abducibles (mental states). It is defined here that when it is said that “model \( M \) reflects the set of rules \( R \)”, this phrase has the above meaning.

In our approach, a relation is imposed between elements derived from some set of rules and models which reflect those rules. This relation is termed the explanatory interpretation of those rules, and acts as a constraint on models. As models are grounded in rules, in order to preserve Property 3 it is assumed that those rules respect the assumption of rationality. Different possible interpretations are presented in the remainder of this paper. Interpretation 1 ensures that the observer ascribes a mental state (belief and goal) to the agent if it has observed a sequence of actions corresponding to the prefix of an observable sequence generated by a plan which relates to aforementioned mental state by means of some known rule.

**Interpretation 1.**

\[
\forall(\gamma \land \beta \land \pi) \in R \forall \delta \in L_{\Delta} \forall \delta' \in \text{OS}(\pi) : \quad (\delta \not< \delta') \implies M \models \text{Obs}(\delta) \rightarrow <\delta'> \rightarrow (G \gamma \land B \beta)
\]

Because it can occur that the agent has not yet finished its plan at the point of explanation, just as that the agent might have dropped its plan before completion, in any state every observed prefix is considered. The above interpretation gives rise to Theorem 1, which makes use of the function \( \text{Obs}(\delta, \pi) = \{ w \mid M, w \models \text{Obs}(\delta) \} \). Of a model \( M \) that reflects a set of rules \( R \) under Interpretation 1 it is said that “\( M \) \text{Int.-n.-reflects } R”.

**Theorem 1.** Let \( M \text{ Int.-1-reflects } R \), then

\[
\forall \delta \in L_{\Delta} \forall w \in \text{OS}(\Delta, \delta) : \{ \gamma, \beta \} \in \text{Obd}(\delta, R) \implies \exists w' \in W : ((w', w) \in \gamma(\delta) \land M, w' \models G \gamma \land B \beta)
\]

Proof. Assume \( \text{Obd}(\delta, R) \neq \emptyset \), and pick some \( (\gamma, \beta) \in \text{Obd}(\delta, R) \). From the definition of \( \text{Obd} \) follows \( \exists(\gamma < \beta \land \pi) \in R \exists \delta' \in \text{OS}(\pi) : (\delta \not< \delta') \) such that \( M \models \text{Obs}(\delta) \rightarrow <\delta'> \rightarrow (G \gamma \land B \beta) \). This implies that \( \forall w \in \text{OS}(\Delta, \delta) : M, w \models <\delta'> \rightarrow (G \gamma \land B \beta) \). Choose any \( w \in \text{OS}(\Delta, \delta) \), such that according to the semantics \( \exists w' \in W : (w', w') \in \gamma(\delta)^{-1} \land w' \in \gamma(\delta) \land w' \models G \gamma \land B \beta \). 

\[
\begin{align*}
\text{Theorem 1 proves a correspondence between the approach of [17] and the models defined in this paper, namely that if for any } \delta \in L_{\Delta} \text{ the function } \text{Obd} \text{ maps to a non-empty set of abducibles (goal/belief pairs), then in each state where } \delta \text{ is observed there exists for each abducible some } \delta \text{-preceding state where this abducible is ascribed. Note that this still allows for significant variety in models, as is shown in Figure 1.}
\end{align*}
\]
depicting models $\mathfrak{M}$ and $\mathfrak{M}'$ which reflect the following rules in different ways (yet both under Interpretation 1).

$p_1 \prec q_1\mid a_1; a_2$ $p_1 \& p_2 \prec q_1\mid a_2; a_3$ $p_3 \prec q_2\mid a_3; a_4$

Assume that, in the respective models, $\{w_5, w_{10}\} \models \text{Obs}(a_1; a_2)$, such that also $\{w_5, w_{10}\} \models \text{Obs}(a_2)$, and furthermore $\{w_2, w_4, w_9\} \models \text{Obs}(a_1)$. Let $\mathfrak{M}, w_1 \models \text{G}p_1 \& \text{B}q_1 \& \neg \text{G}p_2$ whereas $\mathfrak{M}, \{w_3, w_6\} \models \text{G}(p_1 \& p_2) \& \text{B}q_2$. Moreover, $\mathfrak{M}, w_7 \models \text{G}p_2 \& \text{B}q_2$. The right-hand model $\mathfrak{M}'$ is less intricate, seeing that $\mathfrak{M}', w_8 \models \text{G}(p_1 \& p_2) \& \text{B}q_1$ and $\mathfrak{M}', w_9 \models \text{G}p_3 \& \text{B}q_2$.

4.2 Ascription of Plans

In Section 4.1 it was shown through Interpretation 1 how, based on rules, observed sequences relate to ascribed goals and beliefs. In computing observable sequences with the function $\text{OS}$ test actions were discarded, but if test actions which the agent presumably has performed are to be reflected, then computation sequences should be considered instead of observable sequences. Let $CS : L_1 \rightarrow \wp(\mathcal{L}_\Delta)$ be as $CS : L_1 \rightarrow \wp(\mathcal{L}_\Delta)$, except for $CS(\phi) = \{\phi\}^2$. To relate observed sequences to partially executed computation sequences, the relation $\equiv$ is defined on $L_1$ as $\equiv = \{(\pi, \pi), (\pi, \pi'; \pi) | \pi, \pi' \in L_1\}$. The interpretation below is then an alternative to Interpretation 1.

Interpretation 2.

$\forall(\gamma \prec \beta)\in \mathcal{R} \forall \psi \in L_\Delta \forall \psi' \in \text{CS}(\pi):$

$(\exists \pi'' \in L_1 : \pi'' \equiv \pi' \& \text{OS}(\pi'') = \{\delta\}) \implies $

$\mathfrak{M} \models \text{Obs}(\delta) \rightarrow <\pi'' > (G(\gamma) \& B(\beta))$

Proposition 1. Any model $\mathfrak{M}$ which $\text{Int.2-reflects } \mathcal{R}$ also $\text{Int.1-reflects } \mathcal{R}$, such that Theorem 1 applies to $\mathfrak{M}$.

(The proof, which hinges on $(\pi' \in \text{CS}(\pi) \implies \text{OS}(\pi') \subseteq \text{OS}(\pi))$ and $\mathfrak{M}, w \models <\pi'' >\& \text{OS}(\pi'') = \{\delta\} \implies \mathfrak{M}, w \models <\pi'' >\& \text{OS}(\pi'') = \{\delta\}$).

Under Interpretation 2 the possibility is maintained that observed behavior stems from any computation sequence of which the observable prefix matches the observed actions. As Proposition 2 shows, inconclusiveness in ascription cannot be ruled out then, even if it is assumed that ascription based on any single rule in $\mathcal{R}$ does not lead to inconclusiveness.

Proposition 2. Under Interpretation 2 inconclusiveness occurs if $\exists(\gamma \prec \beta)\in \mathcal{R} \exists \pi_1 \in \text{CS}(\pi) \exists \pi_2 \in \mathcal{R}$ This function closely follows $CS$ as in [9], except that it separates elements of computation sequences such that those are in $L_1$.
Proof. Let model $\mathfrak{M}$ Int 2-reflect or Int 3-reflect $\mathcal{R}$, and assume $\exists p < \beta p \to (\sim p < \beta p \to \pi') \in \mathcal{R}$ $\alpha$, $\pi', \pi'' \in \mathcal{L}_g$ ($\pi': \alpha \in CS(\pi') \& \pi'' \\ \alpha \in CS(\pi') \& OS(\pi''; \alpha) = OS(\pi'; \alpha) = (\delta')$). Let $\mathfrak{M}, w = \text{Obs} u, s.t. u \in R$ Int 2 or 3 and Assumption 1 follows $\mathfrak{M}, w = \langle \pi \rangle \langle \beta p \rangle \land \langle \pi'' \rangle (\text{B} \to p), s.t. \exists \mathfrak{M}, w = B\text{Inc} (p)$. Given only knowledge of the agent’s behavior, there is in some cases no way of telling conclusively what the agent should be ascribed, which is illustrated succinctly by the set $\mathcal{R} = \{ \text{on} < \sim \text{on} \mid \text{swtch}, \sim \text{on} < \sim \text{on} \mid \text{swtch} \}$, in a state where switch is observed, and aforementioned constraints cannot enforce that the agent should be conclusively ascribed the belief on or $\sim$ on. The only way then to have an observer draw more “conclusive conclusions” about the agent’s mental state is to use information other than its observed actions.

4.3 Cautious Ascription

In Section 4.2 different interpretations of rules were posited as grounds for plan ascertainment, and it was demonstrated by proof and example that in certain situations inconclusiveness on part of the observer should be admitted. The example in the last paragraph concerning observation of the action switch, which is relevant for achieving both the goals on and $\sim$ on, is in this regard a case in point. It shows that if the observer only knows the rules and actions of the agent, it cannot always be conclusive in ascribing goals or beliefs.

However, in some cases the observer itself has information which allows for stronger claims. Consider the example just presented, and assume that the observer itself is convinced of the fact on, in addition to having observed the agent perform switch and knowing its rules. In this case, the observer can project its belief onto the agent such that B(on) holds. Interpretation 4 reflects this, using the function $\tau : \mathcal{L}_g \to \mathcal{L}$ which translates composed literals $p, \sim p$ to composed literals $\pi, \sim \pi$, and is defined as follows: $\tau (p) = p, \tau (\sim p) = \sim p, \tau (\psi \land \psi') = \tau (\psi) \land \tau (\psi')$, $\tau (\psi \lor \psi') = \tau (\psi) \lor \tau (\psi')$.

**Interpretation 4.** Let $\tau (\psi)$ for any $\psi \in \mathcal{L}_g$, then $\forall (\gamma < \beta \pi) \in \mathcal{R} \forall \pi' \in CS(\pi) \forall \delta \in OS(\pi')$:

$\mathfrak{M}, w = \text{Obs} (\delta) \land \exists \sim \pi' \to (\neg (\gamma) \land \sim \beta) \land [\pi'' \to (\neg (\gamma) \land \sim \beta)] \to [\pi'] (G(\gamma) \land B(\beta))$

In this particular interpretation of behavioral rules, the observer employs its own convictions in ascribing facts as goals or beliefs to the agent. This interpretation is referred to as “cautious ascription” because the observer needs to be convinced of certain facts, other than observed actions, in order to draw conclusions. That the observer makes conclusive ascriptions under this interpretation is proven in Theorem 2, which applies to the class $\mathcal{C}_g$ of models which are constrained by Assumption 1 and Interpretation 4, where $\mathcal{R}$ denotes the set of rules with regard to which both constraints apply.

**Theorem 2.** Let $\tau (\psi)$ for any $\psi \in \mathcal{L}_g$, then $\forall (\gamma < \beta \pi) \in \mathcal{R} \forall \pi' \in CS(\pi) \forall \delta \in OS(\pi')$:

$\mathcal{C}_g = (\text{Obs} (\delta) \land \exists \sim \pi' \to (\neg (\gamma) \land \sim \beta) \land [\pi'' \to (\neg (\gamma) \land \sim \beta)] \to B\text{Con} (\gamma) \land [\pi'' \to G(\gamma) \land B(\beta)]$ 

Proof. Choose any $\mathfrak{M} \in \mathcal{C}_g$ to show that it satisfies the formula mentioned. If the antecedent of $\to$ is false, then it does so by nature of implication. Else, choose any $(\gamma < \beta \pi) \in \mathcal{R}$ and $w \in W$ s.t. $\exists \pi' \in CS(\pi) \exists \delta \in OS(\pi')$ : $w = \text{Obs} (\delta) \land \exists \sim \pi' \to (\neg (\gamma) \land \sim \beta) \land [\pi'' \to (\neg (\gamma) \land \sim \beta)]$. Since $\mathfrak{M}$ is arbitrary and $\mathcal{C}_g$ is determined solely by $\Pi_4$ and $\Pi_1$, at least one of those constraints must ensure ascertainment. Assume $\pi, \pi', \delta, w$ fixed for the remainder of the proof. First, the proof is established for an arbitrary $p, q \in \text{Atom}$ by proof from contradiction (denoted $\perp$). Let $\beta \equiv p, s.t. from \Pi_4 follows w = \langle \pi'' \rangle [Bp \land \langle \pi'' \rangle (\text{B} \to p)], given this, $w = \langle \pi'' \rangle [B\text{Con} (p)$ is established by attempt to derive $w = \langle \pi'' \rangle [Bp \land \langle \pi'' \rangle (\text{B} \to p)]$. Since $w = \langle \pi'' \rangle [Bp]$ it follows that $w = \langle \pi'' \rangle [\text{B} \to p]$ must hold, on grounds of $\Pi_4 or \Pi_1$. For $\Pi_4 this means $w = \langle \pi'' \rangle [Bp \land \langle \pi'' \rangle (\text{B} \to p)]$, i.e. $w = \langle \pi'' \rangle [\text{B} \to p \land \langle \pi'' \rangle (\text{B} \to p)]$. It is given (noting $\beta \equiv p$) that $w = \langle \pi'' \rangle [\text{B} \to p \land \langle \pi'' \rangle (\text{B} \to p)]$, though, s.t. $\perp$. For $w = \langle \pi'' \rangle [\text{B} \to p]$ to derive from $\Pi_4 means \exists \gamma < \beta \pi \in \mathcal{R} \exists \pi'' \in CS(\pi')$ : $w = \langle \pi'' \rangle [\text{B} \to p \land \langle \pi'' \rangle (\text{B} \to p)]$, where $\{G(\gamma) \equiv \Gamma \sim \pi'$. For this to be the case $G(\gamma) \land B \beta$ must derive from $\Pi_4$, which (since $G(\gamma) \equiv \Gamma \sim \pi'$) implies $w = \langle \pi'' \rangle [\text{B} \to p \land \langle \pi'' \rangle (\text{B} \to p)]$, conflicting again with $w = \langle \pi'' \rangle [\text{B} \to p $ in relation to $\Pi_1$. Assumption of $w = \langle \pi'' \rangle [\text{B} \to p \land \langle \pi'' \rangle (\text{B} \to p)]$ (in which case only $\Pi_4 need be considered), from which then follows $w = \beta \pi \delta$ in relation to $\Pi_1$. The proof for arbitrary $\gamma, \beta \equiv q \in \text{Lit}$ proceeds very similarly to the above. Let $Q \equiv B(\text{Con}, G(\text{Con}))$, and note that from the definition of $Q$ follows $Q(\psi \lor \psi') if Q(\psi) \equiv Q(\psi)$ and $Q(\psi \lor \psi') if Q(\psi) \lor Q(\psi'), such that by structural induction the above holds for $\beta, \gamma \in \mathcal{L}_g$.

To put the proof above in the context of an example, assume $\mathcal{R} = \{ \text{on} < \sim \text{on} \mid \text{swtch}, \sim \text{on} < \sim \text{on} \mid \text{swtch} \}$ once more. Let $\mathfrak{M} \in \mathcal{C}_g$ such that $w = \text{Obs} (\text{swtch}) \land \sim \text{on} < \sim \text{on} \mid \text{swtch} \equiv \exists \text{on} \mid \text{swtch} \equiv \text{on} \mid \text{swtch} \equiv \text{on} \mid \text{swtch} \equiv \text{on}$, then from Theorem 2 follows $w = \beta \text{Con} (\sim \text{on} \mid \text{swtch} \equiv \beta \text{Con} (\sim \text{on} \mid \text{swtch} \equiv \text{on})$. Observe that if, e.g., the observer’s sensors malfunction such that $w' = \text{Obs} (\text{swtch}) \land \sim \text{on} < \sim \text{on} \mid \text{swtch} \equiv \text{on} \mid \text{swtch} \equiv \text{on}$, conclusive ascription is not ensured. More elaborate examples are conceivable, but prohibited by space restrictions.

5 Related Work

This work is situated at the junction of the fields of agent programming, dynamic logic, plan recognition, and user modeling. In each respective field there has been copious work, yet to the best of our knowledge there has been very little work that, like ours, falls within the intersection of those fields. This discussion of related work is focused on work closely related to ours both theoretically and technically; more distantly related work is mentioned at other points in the paper.

Literature on plan recognition abounds, see [3] for an overview. A possible division of this field is into probabilistic approaches [4] and approaches based on logic [11, 2]. For probabilistic approaches to have more expressive power than (classical) logic-based approaches, knowledge of the relative probabilities of actions or goals is required. Given that such information is absent in our domain, a probabilistic approach is not particularly attractive. Another division of this field is into keyhole recognition (where the agent is unaware of it
One logic-based approach that is closely related to ours is that of Appelt & Pollack [2], called weighted abduction. In this approach literals are assigned weights, and an observed fact is explained with respect to a theory by either assuming it or proving its antecedent literals; which option is chosen depends on its calculated cost. The proof procedure is accompanied by a model evaluation function, giving preference to models in a way which reflects the weights employed by the algorithm. A drawback of the approach is that observed actions are mere propositions, whose nature as dynamic entities is not inherent in the semantics as it is with dynamic logic. Furthermore, as we objected against probabilistic approaches, the assignment of weights to literals is not apparent in our domain. Also, the work of Appelt & Pollack [2] is based on the BDI- formalization of Cohen & Levesque [5], whereas most work in agent programming is based on that of Rao & Georgeff [16].

Apart from our own earlier work, dynamic logic is used in the context of plan recognition by Rao [15]. The latter work is similar to ours in the sense that correspondence is shown between plan recognition algorithms and a (dynamic agent) logic. It differs significantly, though, in the fact that focus is on the procedural aspects of plan recognition/executions, and in the fact that concepts such as beliefs and goals (desires) are mentioned informally, but are not part of the formalism.

Our semantics of mentalistic expressions is similar to that of Dragoni et al. [8], who use ‘BDI-atoms’ for belief and intention in mental state recognition. Their multi-agent approach formally is a multi-context system, and interpretation of formulae with nested operators is realized with bridge rules. A major difference with our approach is that they consider speech acts, which inherently carry mentalistic information that in our case is not perceptible and can only be presumed. Moreover, we consider structured intentions through plan ascription and separate ‘intention’ from ‘goal’, whereas the term ‘intention’ in the work of Dragoni et al. appears to have the meaning of the term ‘goal’ in our approach.

6 Conclusion and Future Research

In this paper an account of mental state ascription is given using a formalism based on propositional dynamic logic with converse actions and paraconsistent abduction semantics. This account is based on mental state abduction [17], which consists of inferring, based on a set of behavior-generating rules, a goal/belief pair (mental state) which plausibly explains an agent’s observed actions. Different interpretations are given of this approach, of which Interpretation 1 can be considered the canonical interpretation, for which correspondence is shown between properties of models and the output of the mental state abduction function. It is subsumed by Interpretation 2 which incorporates ascription of plan fragments that relate to observed actions, which is shown to lead to inconclusiveness in some cases. Reformulation in terms of Interpretation 3 resolves inconclusiveness in those cases, but does not do so given the assumption that the agent believes the facts constituting some ascribed goal after completing an appropriate plan. At the cost of requiring more information to allow ascription, Interpretation 4 is shown to give rise to conclusive ascription even if that assumption is made.

Interestingly, mental states whose inference in [17] has an abductive basis, are ascribed through deduction in the explanatory interpretations posited here; the reason for this is that dynamic modalities allow for deducing ‘possible paths to the past’. This property is favorable if some form of automated reasoning is used to implement this approach, which also goes for our state-based ascription semantics. In this respect we consider axiomatizing our approach. If a more fine-grained framework is employed (e.g. a logic with modal semantics for beliefs, goals, and such [5, 16, 10], as opposed to our belief/goal atoms), our approach should provide a good starting point for excursions into this direction. Also, the fact that observation of actions is independent of their (presumed) occurrence opens up the possibility of investigating incomplete observation (see [17]). Last, but not least, integrating accounts of mental state inference based on what is claimed (as in [8]) with those based on what is done (as in this paper), offers a promising course for investigating deliberate deception, which is still relatively uncharted territory.

References