

Mental State Ascription Using Dynamic Logic

Michal Sindlar and Mehdi Dastani and John-Jules Meyer¹

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 which interact with other characters whose internal state they cannot access, either by circumstance or by design. 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 programmed as BDI-based software agents to have access to each other's rule libraries, such that they can (defeasibly) infer others' mental states and act on that information. Explanatory functions are defined which take observed actions and a set of behavioral rules as arguments, and map these 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 [6] agent which follows a non-interleaving execution strategy. The formalism also al-

lows 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 $\gamma \leftarrow \beta \mid \pi$, which state that to achieve goal γ , if β is believed, the plan π is appropriate and can be selected. The BNF grammar of rules is given below.

```

<library> ::= <pgrule>+
<pgrule>  ::= <gq>"<->"<bq>"|"<plan>
<plan>    ::= <action> | <test> | <seqplan> | <ifplan> | <whplan>
<test>    ::= "B("<bq>")" | "G("<gq>")"
<bq>      ::= <literal> | <bq>"and"<bq> | <bq>"or"<bq>
<gq>      ::= <literal> | <gq>"and"<gq>
<seqplan> ::= <plan>";"<plan>
<ifplan>  ::= "if"<test>"then"<plan>"else"<plan>
<whplan>  ::= "while"<test>"do"<plan>

```

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 ψ , is then a basic propositional language with positive and negative literals, composed by conjunction or disjunction.

$$\begin{aligned} \psi &::= p \mid \sim p \mid \top \mid \perp \mid \psi_1 \wedge \psi_2 \mid \psi_1 \vee \psi_2 \\ \pi &::= \alpha \mid \mathbf{B}\psi \mid \mathbf{G}\psi \mid \pi_1; \pi_2 \mid \pi_1 + \pi_2 \mid \pi^* \end{aligned}$$

The operator $?$ represents a test action on propositions of $\mathcal{L}_0$, $;$ represents sequential composition, $+$ non-deterministic choice, and $*$ iteration. In the following it is assumed that for rules of the form $\gamma \leftarrow \beta \mid \pi$ it holds that $\gamma, \beta \in \mathcal{L}_0$ and $\pi \in \mathcal{L}_\Pi$.

¹ {michal,mehdi,jj}@cs.uu.nl (Univ. of Utrecht). This research has been supported by the GATE project, funded by the Netherlands Organization for Scientific Research (NWO) and the Netherlands ICT Research and Innovation Authority (ICT Regie).

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 ψ , the abducible fact ϕ can be (defeasibly) inferred. It is slightly adapted for the particular case of inferring beliefs and goals in relation to plans and observed behavior, as shown below.

<i>from</i>	$\text{goal}(\gamma) \ \& \ \text{belief}(\beta) \implies \text{select}(\pi)$
<i>and</i>	$\text{observed}(\alpha_1 \cdots \alpha_n) \ \& \ \text{related}(\alpha_1 \cdots \alpha_n, \pi)$
<i>infer</i>	$\text{goal}(\gamma) \ \& \ \text{belief}(\beta)$

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 $\delta := \alpha \mid \delta_1; \delta_2$ be the definition of the percept language $\mathcal{L}_\Delta \subseteq \mathcal{L}_\Pi$ with typical element δ . The function $OS : \mathcal{L}_\Pi \rightarrow \wp(\mathcal{L}_\Delta)$ translates complex expressions, which include tests and looping or branching constructs, to a set of sequences of observable actions.

$$\begin{aligned}
OS(\alpha) &= \{\alpha\} \\
OS(\phi?) &= \emptyset \\
OS(\pi_1; \pi_2) &= OS(\pi_1) \bullet OS(\pi_2) \\
OS(\pi_1 + \pi_2) &= OS(\pi_1) \cup OS(\pi_2) \\
OS(\pi^*) &= \bigcup_{n \in \mathbb{N}_1} OS(\pi^n), \text{ where } \pi^1 \triangleq \pi \ \& \ \pi^{n+1} \triangleq \pi; \pi^n
\end{aligned}$$

The composition operator $\bullet : \wp(\mathcal{L}_\Delta) \times \wp(\mathcal{L}_\Delta) \rightarrow \wp(\mathcal{L}_\Delta)$ takes arguments $\Delta_1, \Delta_2 \subseteq \mathcal{L}_\Delta$ and maps them to $\{\delta_1; \delta_2 \mid \delta_1 \in \Delta_1, \delta_2 \in \Delta_2\}$ iff $\Delta_1 \neq \emptyset \ \& \ \Delta_2 \neq \emptyset$, to Δ_1 iff $\Delta_2 = \emptyset$, to Δ_2 iff $\Delta_1 = \emptyset$, or to $\emptyset$ otherwise. Note that π^0 is not considered for π^* , 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 rule 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 $\preceq$ on $\mathcal{L}_\Delta$, defined as $\preceq = \{(\delta, \delta), (\delta, \delta; \delta') \mid \delta, \delta' \in \mathcal{L}_\Delta\}$. The abductive explanatory function **abd**, defined below, takes an action sequence and a set of rules as arguments and maps these to goal/belief pairs if the observed sequence is a prefix of an observable trace of the corresponding plan.

$$\text{abd}(\delta, \mathcal{R}) = \{(\gamma, \beta) \mid \exists(\gamma < - \beta \mid \pi) \in \mathcal{R} \ \exists \delta' \in OS(\pi) : \delta \preceq \delta'\}$$

Observe that the function **abd** has the property $\forall \delta, \delta' \in \mathcal{L}_\Delta : (\delta \preceq \delta' \implies \text{abd}(\delta', \mathcal{R}) \subseteq \text{abd}(\delta, \mathcal{R}))$. 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 $\mathcal{L}$, represented by ϕ , let (given $\mathcal{L}_0, \mathcal{L}_\Pi$ as defined in Section 2) $\psi \in \mathcal{L}_0$ and $\pi \in \mathcal{L}_\Pi$.

$$\phi ::= p \mid \mathbf{B}\psi \mid \mathbf{G}\psi \mid \mathbf{O}\alpha \mid \neg\phi \mid \phi_1 \wedge \phi_2 \mid [\pi]\phi \mid [\pi^-]\phi$$

The language thus defined is a variant of PDL with converse actions and poor tests. Note that the negation sign $\sim$ in $\mathcal{L}_0$ negates only single atoms, whereas $\neg$ ranges over propositions such that from $\neg$ and $\wedge$ other logical connectives are derived.

The language $\mathcal{L}$ serves to describe an observer which observes and explains the actions of some agent. This perspective is important, because the operators **B** and **G** 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** take only propositions $\psi \in \mathcal{L}_0$, and **O** only actions $\alpha \in \text{Act}$. The modalities $[\pi]$ and $[\pi^-]$ allow for qualifying propositions that necessarily hold after and before execution of π , and have as duals $\langle \pi \rangle \triangleq \neg[\pi]\neg$ and $\langle \pi^- \rangle \triangleq \neg[\pi^-]\neg$.

3.2 Semantics

The language $\mathcal{L}$ is interpreted in Kripke models $\mathfrak{M}$, such that $\mathfrak{M} = \langle W, \{R_\alpha \mid \alpha \in \text{Act}\}, \vartheta_B, \vartheta_G, \vartheta_I, \vartheta_O \rangle$ where W is a set of states, $R_\alpha \subseteq W \times W$ the accessibility relation for primitive action α , and the ϑ 's are valuation functions. Let $\text{Lit} = \{p, \sim p \mid p \in \text{Atom}\}$. The valuation functions $\vartheta_B : \text{Lit} \rightarrow \wp(W)$ and $\vartheta_G : \text{Lit} \rightarrow \wp(W)$ assign states to literals, $\vartheta_I : \text{Atom} \rightarrow \wp(W)$ assigns states to atoms, and $\vartheta_O : \text{Act} \rightarrow \wp(W)$ reifies the observation of actions. The function $\vartheta : \mathcal{L} \rightarrow \wp(W)$ is the valuation function for propositions in $\mathcal{L}$ 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 ϑ_Q the corresponding function ϑ_B and ϑ_G , respectively, for which holds $\vartheta_Q(\top) = W$ and $\vartheta_Q(\perp) = \emptyset$.

$$\begin{aligned}
\vartheta(p) &= \vartheta_I(p) & , \vartheta(\neg\phi) &= W - \vartheta(\phi) \\
\vartheta(\phi_1 \wedge \phi_2) &= \vartheta(\phi_1) \cap \vartheta(\phi_2) & , \vartheta(\mathbf{Q}\phi) &= \vartheta_Q(\phi) \\
\vartheta_Q(\psi_1 \wedge \psi_2) &= \vartheta_Q(\psi_1) \cap \vartheta_Q(\psi_2) & , \vartheta(\mathbf{O}\alpha) &= \vartheta_O(\alpha) \\
\vartheta_Q(\psi_1 \vee \psi_2) &= \vartheta_Q(\psi_1) \cup \vartheta_Q(\psi_2) \\
\vartheta([\pi]\phi) &= \{w \mid \forall w' \in W : (w, w') \in \varrho(\pi) \implies w' \in \vartheta(\phi)\} \\
\vartheta([\pi^-]\phi) &= \{w \mid \forall w' \in W : (w, w') \in \varrho(\pi)^{-1} \implies w' \in \vartheta(\phi)\}
\end{aligned}$$

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 $\pi \in \mathcal{L}_\Pi$ is based on a function $\varrho : \mathcal{L}_\Pi \rightarrow \wp(W \times W)$, defined as follows.

$$\begin{aligned}
\varrho(\alpha) &= R_\alpha \\
\varrho(\mathbf{Q}\phi?) &= \{(w, w) \mid w \in \vartheta(\mathbf{Q}\phi)\} \\
\varrho(\pi_1; \pi_2) &= \varrho(\pi_1) \circ \varrho(\pi_2) \\
\varrho(\pi_1 + \pi_2) &= \varrho(\pi_1) \cup \varrho(\pi_2) \\
\varrho(\pi^*) &= \bigcup_{n \in \mathbb{N}_0} \varrho(\pi^n), \text{ where } \pi^0 \triangleq (\text{skip})
\end{aligned}$$

In the definition of ϱ , 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 $\varrho(\pi)$. Zero-length iteration of π is defined as the **skip** action, which in this framework corresponds to $\mathbf{Q}(\top)?$ for $\mathbf{Q} \in \{\mathbf{B}, \mathbf{G}\}$. Given a Kripke model $\mathfrak{M}$, the fact that a state w satisfies a formula ϕ is defined as

$$\mathfrak{M}, w \models \phi \text{ iff } w \in \vartheta(\phi)$$

$\mathfrak{M} \models \phi$ denotes that any state in $\mathfrak{M}$ satisfies ϕ , and $\models \phi$ denotes that ϕ is valid in all of the models considered. If $\mathcal{M}$ is a class of models then $\mathcal{M} \models \phi$ is equivalent to $\forall \mathfrak{M} \in \mathcal{M} : \mathfrak{M} \models \phi$. As a shorthand $\mathfrak{M}, \{w_1, \dots, w_n\} \models \phi$ can be used instead of $\mathfrak{M}, w_1 \models \phi$ and $\dots$ and $\mathfrak{M}, w_n \models \phi$. Also, if $w \models \phi$ is used then $\mathfrak{M}$ is assumed to be evident from context.

3.3 Basic Properties

We assume the proposition $\mathbf{Done}(\alpha) \in \mathbf{Atom}$ for each $\alpha \in \mathbf{Act}$ to be interpreted as “action α has been done” and defined as follows: $\forall \alpha \in \mathbf{Act} \forall w \in W : (\exists w' \in W : (w', w) \in R_\alpha) \iff (w \in \vartheta(\mathbf{Done}(\alpha)))$. Furthermore it is assumed that in any state only one single action is considered ‘done’, as opposed to multiple actions having been done simultaneously. This assumption is reflected in Property 1.

$$(\mathfrak{M}, w \models \mathbf{Done}(\alpha) \wedge \mathbf{Done}(\alpha')) \implies (\alpha = \alpha') \quad (1)$$

Observation of actions is formalized through the valuation function ϑ_O , which assigns actions to states. It is here assumed that an action which is considered ‘observed’ is also considered ‘done’, leading to Property 2.

$$\models \mathbf{O}\alpha \rightarrow \mathbf{Done}(\alpha) \quad (2)$$

In order to be able to talk about sequences of observed actions in this our logical language, the expression $\mathbf{Obs}(\delta)$ for any $\delta \in \mathcal{L}_\Pi$ is defined in terms of primitives.

$$\mathbf{Obs}(\alpha_1; \dots; \alpha_n) \triangleq \mathbf{O}\alpha_n \wedge \langle \alpha_1; \dots; \alpha_n \rangle \top \wedge [\alpha_n^-] \mathbf{O}\alpha_{n-1} \wedge [\alpha_{n-1}^-] \mathbf{O}\alpha_{n-2} \wedge \dots \wedge [\alpha_2^-] \mathbf{O}\alpha_1$$

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.

$$\models \mathbf{G}\psi \rightarrow \neg \mathbf{B}\psi \quad (3)$$

It should be noted that ascription by means of the operators $\mathbf{B}$ and $\mathbf{G}$, as presented in Section 3.2, is paraconsistent [14]. This means that for any $p \in \mathbf{Atom}$ it can be the case that neither p nor its literal negation $\sim p$ 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*.

$$\mathbf{BCon}(p) \triangleq \mathbf{B}p \wedge \neg \mathbf{B}\sim p, \quad \mathbf{BCon}(\sim p) \triangleq \mathbf{B}\sim p \wedge \neg \mathbf{B}p$$

$$\mathbf{BInc}(p) \triangleq \mathbf{B}p \wedge \mathbf{B}\sim p, \quad \text{likewise for } \mathbf{BInc}(\sim p)$$

$$\mathbf{BUnk}(p) \triangleq \neg(\mathbf{B}p \vee \mathbf{B}\sim p), \quad \text{likewise for } \mathbf{BUnk}(\sim p)$$

Replacing $\mathbf{B}$ with $\mathbf{G}$ 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: $\mathbf{Q}(\psi \wedge \psi') \triangleq \mathbf{Q}(\psi) \wedge \mathbf{Q}(\psi')$ and $\mathbf{Q}(\psi \vee \psi') \triangleq \mathbf{Q}(\psi) \vee \mathbf{Q}(\psi')$, where $\mathbf{Q} \in \{\mathbf{BCon}, \mathbf{BInc}, \mathbf{BUnk}, \mathbf{GCon}, \mathbf{GInc}, \mathbf{GUnk}\}$.

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 $\gamma \prec \beta \mid \pi$ can be informally described as “the plan π is appropriate for achieving the goal γ if β 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 $\mathfrak{M}$ reflects the set of rules $\mathcal{R}$ ”, 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.

$$\forall (\gamma \prec \beta \mid \pi) \in \mathcal{R} \forall \delta \in \mathcal{L}_\Delta \forall \delta' \in OS(\pi) : (\delta \preceq \delta') \implies \mathfrak{M} \models \mathbf{Obs}(\delta) \rightarrow \langle \delta^- \rangle (\mathbf{G}\gamma \wedge \mathbf{B}\beta)$$

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 $\mathcal{OBS}(\mathfrak{M}, \delta) = \{w \mid \mathfrak{M}, w \models \mathbf{Obs}(\delta)\}$. Of a model $\mathfrak{M}$ that reflects a set of rules $\mathcal{R}$ under Interpretation n it is said that “ $\mathfrak{M}$ Int. n -reflects $\mathcal{R}$ ”.

Theorem 1. *Let $\mathfrak{M}$ Int.1-reflect $\mathcal{R}$, then*

$$\forall \delta \in \mathcal{L}_\Delta \forall w \in \mathcal{OBS}(\mathfrak{M}, \delta) : (\exists (\gamma, \beta) \in \mathbf{abd}(\delta, \mathcal{R}) \implies \exists w' \in W : ((w', w) \in \varrho(\delta) \ \& \ \mathfrak{M}, w' \models \mathbf{G}\gamma \wedge \mathbf{B}\beta))$$

Proof. Assume $\mathbf{abd}(\delta, \mathcal{R}) \neq \emptyset$, and pick some $(\gamma, \beta) \in \mathbf{abd}(\delta, \mathcal{R})$. From the definition of $\mathbf{abd}$ follows $\exists (\gamma \prec \beta \mid \pi) \in \mathcal{R} \exists \delta' \in OS(\pi) : (\delta \preceq \delta')$ such that $\mathfrak{M} \models \mathbf{Obs}(\delta) \rightarrow \langle \delta^- \rangle (\mathbf{G}\gamma \wedge \mathbf{B}\beta)$. This implies that $\forall w \in \mathcal{OBS}(\mathfrak{M}, \delta) : \mathfrak{M}, w \models \langle \delta^- \rangle (\mathbf{G}\gamma \wedge \mathbf{B}\beta)$. Choose any $w \in \mathcal{OBS}(\mathfrak{M}, \delta)$, such that according to the semantics $\exists w' \in W : (w, w') \in \varrho(\delta)^{-1} \ \& \ w' \in \vartheta(\mathbf{G}\gamma \wedge \mathbf{B}\beta)$, i.e. $\exists w' \in W : (w', w) \in \varrho(\delta) \ \& \ \mathfrak{M}, w' \models \mathbf{G}\gamma \wedge \mathbf{B}\beta$. $\square$

Theorem 1 proves a correspondence between the approach of [17] and the models defined in this paper, namely that if for any $\delta \in \mathcal{L}_\Delta$ the function $\mathbf{abd}$ maps to a non-empty set of abducibles (goal/belief pairs), then in each state where δ is observed there exists for each abducible some δ -preceding state where this abducible is ascribed. Note that this still allows for significant variety in models, as is shown in Figure 1

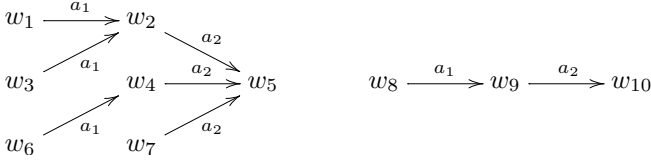


Figure 1. Two models, $\mathfrak{M}$ (left) and $\mathfrak{M}'$ (right).

depicting models $\mathfrak{M}$ and $\mathfrak{M}'$ which reflect the following rules in different ways (yet both under Interpretation 1).

$$p_1 \prec -q_1 \mid a_1; a_2 \quad p_1 \wedge p_2 \prec -q_1 \mid a_1; a_2; a_3 \quad p_3 \prec -q_2 \mid a_2; a_4$$

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

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 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 $CS : \mathcal{L}_\Pi \rightarrow \wp(\mathcal{L}_\Pi)$ be as $OS : \mathcal{L}_\Pi \rightarrow \wp(\mathcal{L}_\Delta)$, except for $CS(\phi?) = \{\phi?\}$.² To relate observed sequences to partially executed computation sequences, the relation $\preceq$ is defined on $\mathcal{L}_\Pi$ as $\preceq = \{(\pi, \pi), (\pi, \pi'; \pi') \mid \pi, \pi' \in \mathcal{L}_\Pi\}$. The interpretation below is then an alternative to Interpretation 1.

Interpretation 2.

$$\begin{aligned} \forall(\gamma \prec -\beta \mid \pi) \in \mathcal{R} \forall \delta \in \mathcal{L}_\Delta \forall \pi' \in CS(\pi) : \\ (\exists \pi'' \in \mathcal{L}_\Pi : \pi'' \preceq \pi' \wedge OS(\pi'') = \{\delta\}) \implies \\ \mathfrak{M} \models \mathbf{Obs}(\delta) \rightarrow \langle \pi''^- \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta)) \end{aligned}$$

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

(The proof, which hinges on $(\pi' \in CS(\pi) \implies OS(\pi') \subseteq OS(\pi))$ and $(\mathfrak{M}, w \models \langle \pi''^- \rangle \top \wedge OS(\pi'') = \{\delta\} \implies \mathfrak{M}, w \models \langle \delta^- \rangle \top)$, is omitted for space conservation.)

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 $\exists(\gamma \prec -\beta \mid \pi), (\gamma' \prec -\beta' \mid \pi') \in \mathcal{R} \exists \pi_1 \in CS(\pi) \exists \pi_2 \in$

$CS(\pi') \exists \pi_3, \pi_4, \mathbf{Q}(\psi)?, \mathbf{Q}(\psi')? \in \mathcal{L}_\Pi : (\pi_3; \mathbf{Q}(\psi)? \preceq \pi_1 \wedge \pi_4; \mathbf{Q}(\psi')? \preceq \pi_2 \wedge OS(\pi_3) = OS(\pi_4) \wedge \exists \psi'' \in \mathcal{L}_0 : (\{\mathbf{Q}(\psi), \mathbf{Q}(\psi')\} \models \mathbf{QInc}(\psi'')))$, where $\mathbf{Q} \in \{\mathbf{B}, \mathbf{G}\}$.

Proof. Let $\mathcal{R} = \{p_1 \prec -q_1 \mid \pi_1, p_2 \prec -q_2 \mid \pi_2\}$, where $\pi_1 = a_1; \mathbf{B}(r)?; a_2$ and $\pi_2 = a_1; \mathbf{B}(\sim r)?; a_3$, such that $CS(\pi_1) = \{\pi_1\}$ and $CS(\pi_2) = \{\pi_2\}$. Let $\mathfrak{M}$ Int.2-reflect $\mathcal{R}$. It holds for $\pi'_1 = a_1; \mathbf{B}(r)?$ and $\pi'_2 = a_1; \mathbf{B}(\sim r)?$ that $\pi'_1 \preceq \pi_1$ and $\pi'_2 \preceq \pi_2$, and $OS(\pi'_1) = OS(\pi'_2) = \{a_1\}$. Let $\mathfrak{M}, w \models \mathbf{O}a_1$ s.t. $\mathfrak{M}, w \models \langle \pi'_1^- \rangle \top \wedge \langle \pi'_2^- \rangle \top$ follows from Int. 2. From $\mathfrak{M}, w \models \langle a_1; \mathbf{B}(r)?^- \rangle \top \wedge \langle a_1; \mathbf{B}(\sim r)?^- \rangle \top$ follows $\mathfrak{M}, w \models \mathbf{B}(r) \wedge \mathbf{B}(\sim r)$ and therefore $\mathfrak{M}, w \models \mathbf{BInc}(r)$. $\square$

The reason that inconclusiveness occurs in some cases, as shown in Proposition 2, is that Interpretation 2 considers *all* (prefixes of) computation sequences whose actions match the observed actions. This includes prefixes of computation sequences which end with test actions on propositions, and because Interpretation 2 demands that for any such prefix π'' there exists a ‘way back’ (meaning $\langle \pi''^- \rangle \top$) it must be the case that test actions at the end of the computation sequence π'' succeed in the state in which the observable sequence $\delta \in OS(\pi'')$ is observed. Conceptually, it seems more appropriate to consider only test actions which have (presumably) been performed *up to the last observed action*. Interpretation 3 is a reformulation of Interpretation 2 under which inconclusiveness need not occur as in Proposition 2.

Interpretation 3.

$$\begin{aligned} \forall(\gamma \prec -\beta \mid \pi) \in \mathcal{R} \forall \delta \in \mathcal{L}_\Delta \forall \pi' \in CS(\pi) : \\ (\exists \pi'' \in \mathcal{L}_\Pi : (\pi'' \preceq \pi' \wedge OS(\pi'') = \{\delta\} \wedge \\ \forall \pi''' \in \mathcal{L}_\Pi : (\pi''' \preceq \pi' \wedge OS(\pi''') = \{\delta\}) \implies (\pi'' \preceq \pi''')))) \implies \\ \mathfrak{M} \models \mathbf{Obs}(\delta) \rightarrow \langle \pi''^- \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta)) \end{aligned}$$

Proposition 3. Inconclusiveness in ascription as occurring in Proposition 2 is avoidable under Interpretation 3.

Proof. Let $\mathfrak{M}$ Int.3-reflect $\mathcal{R}$ and let $\pi', \pi'' \in \mathcal{L}_\Pi$ be such that $\exists(\gamma \prec -\beta \mid \pi) \in \mathcal{R} : \pi' \in CS(\pi) \wedge \pi'' \preceq \pi' \wedge OS(\pi'') = \{\delta\}$. Furthermore, let $\mathfrak{M}, w \models \mathbf{Obs}(\delta)$. If $\exists \pi''', \phi? \in \mathcal{L}_\Pi : \pi'' = \pi''' ; \phi?$ then $\exists \pi'''' \in \mathcal{L}_\Pi : \pi'''' \preceq \pi'' \wedge OS(\pi''') = \{\delta\} \wedge \pi'' \not\preceq \pi''''$ (namely $\pi'''' = \pi''$ or another strict prefix). On the other hand, if $\exists \pi''', \alpha \in \mathcal{L}_\Pi : \pi'' = \pi''' ; \alpha$ then $\forall \pi'''' \in \mathcal{L}_\Pi : (\pi'''' \preceq \pi'' \wedge OS(\pi''') = \{\delta\}) \implies (\pi'' = \pi''')$, such that the π'' of Interpretation 3 must end with a primitive action. By choosing R_α for actions from different sequences to lead to different preceding states, inconclusiveness is avoided. $\square$

Proposition 3 shows that inconclusiveness is avoidable in some cases; however, in other cases it is not (unless additional restrictions are imposed on rules). Consider Assumption 1, which states that if an agent is ascribed a goal/belief pair in some state, then in all states reachable through a plan that is related to that pair by some rule, the agent is ascribed the belief that its former goal holds in those states.

Assumption 1.

$$\forall(\gamma \prec -\beta \mid \pi) \in \mathcal{R} : \mathfrak{M} \models (\mathbf{G}\gamma \wedge \mathbf{B}\beta) \rightarrow [\pi] \mathbf{B}\gamma$$

Proposition 4. Under both Interpretations 2 and 3 inconclusiveness can arise if Assumption 1 is made.

² This function closely follows CS as in [9], except that ; separates elements of computation sequences such that those are in $\mathcal{L}_\Pi$.

Proof. Let model $\mathfrak{M}$ Int.2-reflect or Int.3-reflect $\mathcal{R}$, and assume $\exists(p \leftarrow \beta \mid \pi), (\sim p \leftarrow \beta' \mid \pi') \in \mathcal{R} \exists \alpha, \pi'', \pi''' \in \mathcal{L}_\Pi : (\pi''; \alpha \in CS(\pi) \& \pi'''; \alpha \in CS(\pi') \& OS(\pi''); \alpha) = OS(\pi'''; \alpha) = \{\delta\}$. Let $\mathfrak{M}, w \models \mathbf{Obs}(\delta)$, s.t. under Int. 2 or 3 and Assumption 1 follows $\mathfrak{M}, w \models \langle \pi''; \alpha^- \rangle [\pi](\mathbf{B}p) \wedge \langle \pi'''; \alpha^- \rangle [\pi'](\mathbf{B}\sim p)$, s.t. $\mathfrak{M}, w \models \mathbf{BInc}(p)$. $\square$

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 $\mathcal{R} = \{\text{on} \leftarrow \sim \text{on} \mid \text{switch}, \sim \text{on} \leftarrow \text{on} \mid \text{switch}\}$. In a state where **switch** is observed, aforementioned constraints cannot enforce that the agent should be conclusively ascribed the belief **on** or $\sim \text{on}$. 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 ascription, 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 $\sim \text{on}$, 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 also $\mathbf{B}(\text{on})$ holds. Interpretation 4 reflects this, using the function $\tau : \mathcal{L}_0 \rightarrow \mathcal{L}$ which translates composed literals $p, \sim p$ to composed literals $p, \neg p$, and is defined as follows: $\tau(p) = p$, $\tau(\sim p) = \neg p$, $\tau(\psi \wedge \psi') = \tau(\psi) \wedge \tau(\psi')$, $\tau(\psi \vee \psi') = \tau(\psi) \vee \tau(\psi')$.

Interpretation 4. Let $\hat{\psi} \equiv \tau(\psi)$ for any $\psi \in \mathcal{L}_0$, then

$$\forall(\gamma \leftarrow \beta \mid \pi) \in \mathcal{R} \forall \pi' \in CS(\pi) \forall \delta \in OS(\pi') :$$

$$\mathfrak{M} \models \mathbf{Obs}(\delta) \wedge \hat{\gamma} \wedge \langle \pi'^- \rangle (\neg(\hat{\gamma}) \wedge \hat{\beta}) \wedge [\pi'^-](\neg(\hat{\gamma}) \wedge \hat{\beta}) \rightarrow [\pi'^-](\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))$$

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 $\mathcal{C}^\mathcal{R}$ 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 $\hat{\psi} \equiv \tau(\psi)$ for any $\psi \in \mathcal{L}_0$, then

$$\forall(\gamma \leftarrow \beta \mid \pi) \in \mathcal{R} \forall \pi' \in CS(\pi) \forall \delta \in OS(\pi') :$$

$$\mathcal{C}^\mathcal{R} \models (\mathbf{Obs}(\delta) \wedge \hat{\gamma} \wedge \langle \pi'^- \rangle (\neg(\hat{\gamma}) \wedge \hat{\beta}) \wedge [\pi'^-](\neg(\hat{\gamma}) \wedge \hat{\beta})) \rightarrow \mathbf{BCon}(\gamma) \wedge [\pi'^-](\mathbf{GCon}(\gamma) \wedge \mathbf{BCon}(\beta))$$

Proof. Choose any $\mathfrak{M} \in \mathcal{C}^\mathcal{R}$ to show that it satisfies the formula mentioned. If the antecedent of $\rightarrow$ is false, then it does so by nature of implication. Else, choose any $(\gamma \leftarrow \beta \mid \pi) \in \mathcal{R}$ and $w \in \mathbf{W}$ s.t. $\exists \pi' \in CS(\pi) \exists \delta \in OS(\pi') : (w \models \mathbf{Obs}(\delta) \wedge \hat{\gamma} \wedge \langle \pi'^- \rangle (\neg(\hat{\gamma}) \wedge \hat{\beta}) \wedge [\pi'^-](\neg(\hat{\gamma}) \wedge \hat{\beta}))$. Since $\mathfrak{M}$ is arbitrary and $\mathcal{C}^\mathcal{R}$ is determined solely by I4 and A1, at least one of those constraints must ensure ascription. Assume π, π', δ, w fixed for the remainder of the proof. First, the proof is established for an arbitrary $p, q \in \mathbf{Atom}$ by proof from contradiction (denoted ζ). Let $\beta \equiv p$, s.t. from I4 follows $w \models [\pi'^-]\mathbf{B}p$ and $w \models \langle \pi'^- \rangle \mathbf{B}p$. Given this, $w \models [\pi'^-]\mathbf{BCon}(p)$ is established by attempt to derive $w \models \langle \pi'^- \rangle \neg \mathbf{BCon}(p)$, i.e. $w \models \langle \pi'^- \rangle (\neg \mathbf{B}p \vee \mathbf{B}\sim p)$. Since $w \models [\pi'^-]\mathbf{B}p$ it follows that $w \models \langle \pi'^- \rangle \mathbf{B}\sim p$ must hold, on grounds of I4 or A1. For I4 this means $w \models \langle \pi'^- \rangle \tau(\sim p) \wedge [\pi'^-]\tau(\sim p)$, i.e. $w \models \langle \pi'^- \rangle (\neg p) \wedge [\pi'^-](\neg p)$. It is given (noting $\hat{\beta} \equiv p$) that $w \models \langle \pi'^- \rangle (p) \wedge [\pi'^-](p)$, though, s.t. ζ . For $w \models \langle \pi'^- \rangle (\mathbf{B}\sim p)$ to derive from A1 means $\exists(\gamma'' \leftarrow \beta'' \mid \pi'') \in \mathcal{R} \exists \pi''' \in CS(\pi'') : w \models \langle \pi'''; \pi'^- \rangle (\mathbf{G}\gamma'' \wedge \mathbf{B}\beta'')$, where $\{\mathbf{G}\gamma''\} \models \mathbf{G}\sim p$. For this to be the case $\mathbf{G}\gamma'' \wedge \mathbf{B}\beta''$ must derive from I4, which (since $\mathbf{G}\gamma''$ entails $\mathbf{G}\sim p$) implies $w \models \langle \pi'^- \rangle \tau(\sim p)$, i.e. $w \models \langle \pi'^- \rangle \neg p$, conflicting again with $w \models [\pi'^-]p$, s.t. ζ . Let $\gamma \equiv q$, then $w \models [\pi'^-]\mathbf{GCon}(q)$ is proven likewise by assuming $w \models \langle \pi'^- \rangle \neg \mathbf{GCon}(q)$ (in which case only I4 need be considered), from which then follows $w \models \mathbf{B}q$ in relation to A1. Assumption of $w \models \mathbf{B}\sim q$ leads to ζ in relation to I4, s.t. $w \models \mathbf{BCon}(q)$. The proof for arbitrary $\sim p, \sim q \in \mathbf{Lit}$ proceeds very similarly to the above. Let $\mathbf{Q} \in \{\mathbf{BCon}, \mathbf{GCon}\}$, and note that from the definition of $\mathbf{Q}$ follows $\mathbf{Q}(\psi \wedge \psi')$ if $\mathbf{Q}(\psi) \wedge \mathbf{Q}(\psi')$ and $\mathbf{Q}(\psi \vee \psi')$ if $\mathbf{Q}(\psi) \vee \mathbf{Q}(\psi')$, such that by structural induction the above holds for any $\beta, \gamma \in \mathcal{L}_0$. $\square$

To put the above in the context of an example, assume $\mathcal{R} = \{\text{on} \leftarrow \sim \text{on} \mid \text{switch}, \sim \text{on} \leftarrow \text{on} \mid \text{switch}\}$ once more. Let $\mathfrak{M} \in \mathcal{C}^\mathcal{R}$ such that $w \models \mathbf{O}(\text{switch}) \wedge \neg \text{on} \wedge \langle \text{switch}^- \rangle (\text{on}) \wedge [\text{switch}^-](\text{on})$, then from Theorem 2 follows $w \models \mathbf{BCon}(\sim \text{on}) \wedge [\text{switch}^-](\mathbf{GCon}(\sim \text{on}) \wedge \mathbf{BCon}(\text{on}))$. Observe that if, e.g., the observer's sensors malfunction such that $w' \models \mathbf{O}(\text{switch}) \wedge \neg \text{on} \wedge \langle \text{switch}^- \rangle (\neg \text{on}) \wedge [\text{switch}^-](\neg \text{on})$, 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

being observed) and intended recognition (where the agent structures its activities to facilitate recognition) [3]; our approach is of the former kind.

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/execution, 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 ascription 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

- [1] Natasha Alechina, Mehdi Dastani, Brian Logan, and John-Jules Meyer, ‘A logic of agent programs’, in *Proc. of the Twenty-Second AAAI Conference on AI*, pp. 795–800, (2007).
- [2] Douglas E. Appelt and Martha E. Pollack, ‘Weighted abduction for plan ascription’, *User Modeling and User-Adapted Interaction*, **2**(1-2), 1–25, (1991).
- [3] Sandra Carberry, ‘Techniques for plan recognition’, *User Modeling and User-Adapted Interaction*, **11**(1-2), 31–48, (2001).
- [4] Eugene Charniak and Robert P. Goldman, ‘A Bayesian model of plan recognition’, *Artificial Intelligence*, **64**, 53–79, (1993).
- [5] Philip R. Cohen and Hector J. Levesque, ‘Intention is choice with commitment’, *Artificial Intelligence*, **42**(2-3), 213–261, (1990).
- [6] Mehdi Dastani, ‘2APL: A practical agent programming language’, *Autonomous Agents and Multi-Agent Systems*, **16**, 214–248, (2008).
- [7] Daniel Dennett, *The Intentional Stance*, MIT Press, 1987.
- [8] Aldo Franco Dragoni, Paolo Giorgini, and Luciano Serafini, ‘Mental states recognition from communication’, *Journal of Logic and Computation*, **12**(1), 119–136, (2002).
- [9] David Harel, Dexter Kozen, and Jerzy Tiuryn, *Dynamic Logic*, MIT Press, 2000.
- [10] Wiebe van der Hoek, Bernd van Linder, and John-Jules Meyer, *An Integrated Modal Approach to Rational Agents*, 133–168, Dordrecht: Kluwer, 1998.
- [11] Henry A. Kautz, *A Formal Theory of Plan Recognition and its Implementation*, 69–126, Morgan Kaufmann, 1993.
- [12] John Laird, ‘It knows what you’re going to do: Adding anticipation to a Quakebot’, in *AGENTS*, (2001).
- [13] Shaun Nichols and Stephen P. Stich, *Mindreading*, Oxford University Press, 2003.
- [14] Stanford Encyclopedia of Philosophy: *Paraconsistent Logic*. <http://plato.stanford.edu/entries/logic-paraconsistent/>.
- [15] Anand S. Rao, ‘Means-end plan recognition — towards a theory of reactive recognition’, Technical Report 49, Australian AI Institute, (June 1994).
- [16] Anand S. Rao and Michael P. Georgeff, ‘Modeling rational agents within a BDI-architecture’, in *Proc. of the Second Int. Knowl. Repr. Conf. (KR’91)*, pp. 473–484, (1991).
- [17] Michal Sindlar, Mehdi Dastani, Frank Dignum, and John-Jules Meyer, ‘Mental state abduction of BDI-based agents’, in *Proceedings of DALI@AAMAS’08*, pp. 161–178, (2008).
- [18] Michal Sindlar, Mehdi Dastani, and John-Jules Meyer, ‘BDI-based development of virtual characters with a theory of mind’, in *Proceedings of the Ninth International Conference on Intelligent Virtual Agents*, pp. 34–41, (2009).