A Modal Logic for Beliefs and Pro Attitudes

Kaile Su[1,2] and **Abdul Sattar**[1] and **Han Lin**[3] and **Mark Reynolds**[4]

[1]Institute for Integrated and Intelligent Systems, Griffith University, Brisbane, Australia

[2]Key Laboratory of High Confidence Software Technologies of Ministry of Education

School of Electronics Engineering and Computer Science, Peking University, Beijing, China

[3]Department of Computer Science, Sun Yat-Sen University, Guangzhou, China

[4] School of Computer Science and Software Engineering, The University of Western Australia, Perth, Australia

{k.su, a.sattar}@griffith.edu.au; linhan095@sohu.com; mark@csse.uwa.edu.au

Abstract

Agents' pro attitudes such as goals, intentions, desires, wishes, and judgements of satisfactoriness play an important role in how agents act rationally. To provide a natural and satisfying formalization of these attitudes is a longstanding problem in the community of agent theory. Most of existing modal logic approaches are based on Kripke structures and have to face the so-called side-effect problem. This paper presents a new modal logic formalizing agents' pro attitudes, based on neighborhood models. There are three distinguishing features of this logic. Firstly, this logic naturally satisfies *Bratman's requirements* for agents' beliefs and pro attitudes, as well as some interesting properties that have not been discussed before. Secondly, we give a sound and complete axiom system for characterizing all the valid properties of beliefs and pro attitudes. We introduce for the first time the notion of *linear neighborhood frame* for obtaining the semantic model, and this brings a new member to the family of non-normal modal logics. Finally, we argue that the present logic satisfies an important requirement proposed from the viewpoint of computation, that is, *computational grounding*, which means that properties in this logic can be given an interpretation in terms of some concrete computational model. Indeed, the presented neighborhood frame can be naturally derived from probabilistic programming with utilities.

Introduction

The present work belongs to the subject of formalizing mental attitudes of a rational agent. Agents' mental attitudes include informational attitudes such as knowledge and belief, and pro attitudes such as goal, intention, desire, wish, and judgement of satisfactoriness. Both information and pro attitudes play an important role in how agents act rationally. It is very successful and satisfying to formalize agents' information attitudes by using epistemic logics (Fagin *et al.* 1995). These modal logics for formalizing knowledge and belief have at least two salient points:

- The sound and complete axiom systems of epistemic logics capture all valid properties of knowledge and belief naturally and succinctly.

- The *interpreted system* model (Halpern & Zuck 1992; Halpern & Vardi 1986) offers a natural interpretation, in

terms of the states of computer processes, to S5 epistemic logic, so that formulas in epistemic logic that are valid with respect to the interpreted system can be understood as valid properties of program computations. In this sense, the interpreted system model is *computationally grounded* (Wooldridge 2000).

To provide a natural and satisfying formalization of pro attitudes is a longstanding problem in the community of agent theory. Compared to epistemic logics for formalizing information attitudes, existing modal logic approaches for dealing with pro attitudes are not so successful and satisfying. They are usually based on normal modal logics, so that the pro attitudes characterized by them have to be closed under logical consequences. This is the so-called side-effect problem (Bratman 1987). The well-known theory of intention (Cohen & Levesque 1990) and the formalism of the belief-desire-intention paradigm (Rao & Georgeff 1998), for example, are along this line. Though there are several strategies to make these logics side-effect free, they are still not fully side-effect free (Konolige & Pollack 1993). For instance, these logics accept that desiring p and desiring q imply desiring p or q, which seems reasonable, but, as we will demonstrate in the next section, may surprisingly cause a problem.

Another problem of existing modal logic approaches is that they are generally ungrounded (van der Hoek & Wooldridge 2003). To make a logic of belief, desire and intention computationally grounded is a challenging problem as announced by M. Wooldridge at ICMAS-2000.

This paper presents a new non-normal modal logic formalizing agents' beliefs and pro attitudes. The distinguishing features of this logic are on three aspects. Firstly, this logic naturally satisfies *Bratman's requirements* for agents' beliefs and pro attitudes, as well as some interesting properties that have not been investigated in the literature. Secondly, we give a sound and complete axiom system for characterizing all the valid properties of an agent's beliefs and pro attitudes in some practically interesting cases. We introduce for the first time the notion of *linear neighborhood frame* for obtaining the semantic model. This brings a new member to the family of non-normal modal logics, while most existing modal logics of agency have either direct correspondences in the family of modal logics, or can be regarded as combinations of several existing modal logics. Finally, we argue that the present logic satisfies an important

requirement proposed from the viewpoint of computation, that is, *computational grounding* (Wooldridge 2000), which means that properties in this logic can be given an interpretation in terms of some concrete computational model. Indeed, the present neighborhood model can be naturally derived from the so-called probabilistic programs with utilities[1].

In this logic, we deliberately consider only one agent and only static properties of pro attitudes and do not distinguish different kinds of pro attitudes. However, this logic significantly differs from others even at this level of abstraction.

The remainder of this paper is organized as follows. In the next section, we discuss some desired properties of a modal logic of pro attitudes. Then, we define a new semantic model for beliefs and pro attitudes, with respect to which a sound and complete axiom system is presented. We also demonstrate how our abstract semantic model can be derived from a concrete computational one, that is, probabilistic programming with utilities. Finally, we discuss the related work and conclude the paper.

Desired Properties of a Modal Logic of Pro Attitudes

We begin with discussing some desired properties of a modal logic of pro attitudes from three different aspects, i. e. , philosophy, logic and computation.

Let $\mathbf{At}$ be a set of atomic propositions. Our modal language, denoted by $\mathcal{L}^{B,H}(\mathbf{At})$, is defined as follows:

$$p \mid \neg\varphi \mid \varphi \wedge \varphi \mid B\varphi \mid H\varphi \mid$$

where $p \in \mathbf{At}$. The intended interpretation of $B\varphi$ is "φ is believed", while H is a modality to represent pro attitudes such as "preference","wish", "hope","desire" etc., and $H\varphi$ can be informally understood as " the agent is **h**appy if being told φ", or "φ is good news to the agent". As usual, propositional connectives "$\vee$" (or), "$\Rightarrow$" (imply) and "$\Leftrightarrow$" (if and only if) can be defined by "$\neg$" and "$\wedge$".

From the Viewpoint of Philosophy

The philosophical work of (Bratman 1987) proposed some important properties of agents' mental attitudes such as beliefs, desires and intentions. These properties became very influential on formalizations of agents' mental attitudes. We extend these properties to general pro attitudes and express them in the form of $\models \varphi$ or $\not\models \varphi$, where φ is a formula schema in $\mathcal{L}^{B,H}(\mathbf{At})$. Statement $\models \varphi$ indicates that φ is valid or always true, while $\not\models \varphi$ means that φ is not necessarily true.

- Asymmetry Thesis: $\models H\varphi \Rightarrow \neg B\neg\varphi$ and $\not\models H\varphi \Rightarrow B\varphi$.

  The former says that it is impossible (or irrational) for an agent to desire φ but believe $\neg\varphi$; while the latter means that a rational agent may have incomplete beliefs about her desires.

---

[1]A probabilistic program with utilities here means a probabilistic agent program where a utility function U is defined so that, for each run r of the program, the agent is awarded interest $U(r)$.

- Side-Effect-Free Principle: $\not\models H\varphi \wedge B(\varphi \Rightarrow \psi) \Rightarrow H\psi$.

  This Principle indicates that an agent, who desires φ and believes that φ implies ψ, may not desire ψ

- Non-Transference Principle: $\models B\varphi \Rightarrow \neg H\varphi$.

  It means that if an agent believes some claim φ already, then φ should not be good news to her.

- Introspectivity Assumption: $\models H\varphi \Rightarrow BH\varphi$ and $\models \neg H\varphi \Rightarrow B\neg H\varphi$.

  It indicates that an agent is conscious of her pro attitudes.

Notice that some properties we capture may be stronger than those formalized in (Rao & Georgeff 1998). For instance, the Non-Transference Principle may be presented in a weaker version, say, $\not\models B\varphi \Rightarrow H\varphi$, which seems more appropriate for those pro attitudes like intentions. The correctness of these properties in general is very controversial and we do not intend to take sides in this issue.

We are particularly interested in the Side-Effect-Free Principle. The above version of this principle is weaker and satisfied by Rao and Georgeff's account. We propose some stronger versions of the principle. Consider the case where the "side-effect" of an agent's desire φ is simply a logical consequence of φ, say, $\varphi \vee \psi$. We then get a stronger version of the principle, $\not\models H\varphi \Rightarrow H(\varphi \vee \psi)$. We even propose the following version of the Side-Effect-Free Principle: $\not\models H\varphi \wedge H\psi \Rightarrow H(\varphi \vee \psi)$.

At the first glance, the above version of the Side-Effect-Free Principle seems extremely problematic, because if both φ and ψ are good news to you, then, by your intuition, $\varphi \vee \psi$ should be. However, let us consider the following scenario, where Alice, Bob and Peter apply to the same lectureship position and are all on the short list for review. It would be good news to Alice that Bob fails, because she can then expect more chances to get the position. Also, Alice would be happy being told that Peter fails. However, Alice will never wish that Bob fails or Peter fails, because at most one of them can win and it is inevitably true that Bob fails or Peter fails. This scenario illustrates that an agent with desires φ and ψ should not be forced to have a desire $\varphi \vee \psi$. Nevertheless, in the logic presented in this paper, we have that

$$\models H\varphi \wedge H\psi \wedge B\neg(\varphi \wedge \psi) \Rightarrow H(\varphi \vee \psi).$$

From the Viewpoint of Logic and Computation

As a modal logic, a theory of pro attitudes should have a formal semantic model, with respect to which a sound and complete axiom system could be developed. Furthermore, it is satisfactory that the related decidability and complexity results can be carried out.

From the viewpoint of computer science, (Wooldridge 2000) proposed a requirement for agent theories, namely, computational grounding. A theory of agency is said to be computationally grounded if we can give the theory an interpretation in terms of some concrete computational model. If logics of rational agency are to be taken seriously as specification languages for agent or multi-agent systems, then the requirement of computational grounding is important for them. Moreover, we argue that logics of agency with some concrete computational models are more reliable, because

one might make some purely artificial models to capture some intuitively valid but problematic properties of agency (see also the argument below Proposition 9).

Most logical theories of rational agency, based on possible worlds semantics are generally ungrounded, because there is usually no direct relationship between the accessibility relations that are used to characterize an agent's mental state, and any concrete computational model. It is an open issue to give a computationally grounded semantics to goals (Wooldridge 2000), and the present work takes a significant step towards this challenging issue.

The Neighborhood Model for Pro Attitudes

Neighborhood semantics, also known as Scott-Montague semantics (Scott ; Montague 1970), is a formal semantics for modal logics. It is a generalization, developed independently by Dana Scott and Richard Montague, of the well-known relational semantics for modal logic. Neighborhood semantics is based on neighborhood frames instead of Kripke structures.

The frames considered in this paper are the hybrid of neighborhood frame and Kripke structure. Although the accessibility relation for beliefs in the frames can be represented by a neighborhood function in neighborhood semantics, we feel it is more convenient to keep it as a relation.

Definition 1 A tuple $\langle W, R, N \rangle$ is called a hybrid neighborhood frame, if W is a non-empty set, R is a relation on W and N is a function: $W \to 2^{2^W}$, which is called a neighborhood function.

In the above definition, we denote $\{w' \in W \mid wRw'\}$ by $R[w]$. Intuitively, W is the set of all possible states or worlds, and $R[w]$ is the set of those worlds that are possible or believable to the agent at the current world w.

Definition 2 A hybrid neighborhood model is a tuple $\mathbb{M} = \langle W, R, N, V \rangle$, where $\mathbb{F} = \langle W, R, N \rangle$ is a hybrid neighborhood frame and $V : \mathbf{At} \to 2^W$ is a valuation function.

For short, we use frequently "frame (model)" instead of "hybrid neighborhood frame (model)". Given a model $\mathbb{M} = \langle W, R, N, V \rangle$ and $w \in W$, the truth of a formula φ is defined inductively as follows.

- $\mathbb{M}, w \models p$ iff $w \in V(p)$.
- $\mathbb{M}, w \models \neg\varphi$ iff $\mathbb{M}, w \not\models \varphi$.
- $\mathbb{M}, w \models \varphi \wedge \psi$ iff $\mathbb{M}, w \models \varphi$ and $\mathbb{M}, w \models \psi$.
- $\mathbb{M}, w \models B\varphi$ iff $\mathbb{M}, w' \models \varphi$ for all $w' \in W$ such that wRw'.
- $\mathbb{M}, w \models H\varphi$ iff $(\varphi)^{\mathbb{M}} \in N(w)$, where $(\varphi)^{\mathbb{M}}$ denotes the set $\{w \mid \mathbb{M}, w \models \varphi\}$, called the *truth set* of φ.

As usual, $\mathbb{F} \models \varphi$ stands for that, for all models $\mathbb{M}$ based on frame $\mathbb{F}$ and for all states w, $\mathbb{M}, w \models \varphi$ holds.

The general notion of hybrid neighborhood frames is not appropriate for characterizing beliefs and pro attitudes of a rational agent. Therefore, we propose several specific classes of hybrid neighborhood frames (models).

Definition 3 Given a neighborhood function $N : W \to 2^{2^W}$.

- We say that N is *linear* if, for all $w \in W$ and all $X_1, X_2 \subseteq W$, we have the following:
  1. If $X_1, X_2 \in N(w)$ and $X_1 \cap X_2 = \{\}$, then $X_1 \cup X_2 \in N(w)$.
  2. If $X_1, X_2 \notin N(w)$ and $X_1 \cap X_2 = \{\}$, then $X_1 \cup X_2 \notin N(w)$.
- Neighborhood function N is *unit-zero-exclusive*, if for every $w \in W$, both $W \notin N(w)$ and $\{\} \notin N(w)$ hold .

A frame (model) is called *linear or unit-zero-exclusive*, if the neighborhood function is linear or unit-zero-exclusive, respectively.

Notice that the underlying concrete computation model for pro attitudes here is probabilistic programming with utilities, and the notion of linear frame captures the additivity (or linearity) of probability and utility expectation functions applying a union of two disjoint sets. The notion of unit-zero-exclusion is to characterize those agents who are *realistic* in the sense that they do not desire anything they think of as inevitably true or false (i. e. , $\models \neg H true \wedge \neg H \neg true$ where $true$ is a valid proposition, say $p \vee \neg p$).

For every linear and unit-zero-exclusive neighborhood function N, we have that, for all $w \in W$ and all $X \subseteq W$, $X \notin N(w)$ or $W - X \notin N(w)$. As a result, for a linear and unit-zero-exclusive frame $\mathbb{F}$, we have that $\mathbb{F} \models \neg(H\varphi \wedge H\neg\varphi)$.

Definition 4 A frame $\langle W, R, N \rangle$ is *introspective* if,

1. for all $w, w' \in W$ with wRw', we have that $R[w] = R[w']$ and $N(w) = N(w')$;
2. for all $w \in W$ and all $X, X' \subseteq W$ with $X \cap R[w] = X' \cap R[w]$, we have that $X \in N(w)$ iff $X' \in N(w)$.

Intuitively, the first part of above definition says that the agent is introspective to her own beliefs, while the second indicates that the agent is conscious of her pro attitudes. A model is called *introspective* if the corresponding frame is introspective.

Proposition 5 *For all unit-zero-exclusive and introspective frames $\mathbb{F}$, we have that*

1. $\mathbb{F} \models H\varphi \Rightarrow (\neg B\neg\varphi \wedge \neg B\varphi)$
2. $\mathbb{F} \models (H\varphi \Rightarrow BH\varphi) \wedge (\neg H\varphi \Rightarrow B\neg H\varphi)$

The above proposition indicates that the Asymmetry Thesis, the Non-Transference Principle and the Introspectivity Assumption hold for unit-zero-exclusive and introspective frames.

Proposition 6 *For all linear and introspective frames $\mathbb{F}$, we have that*

1. $\mathbb{F} \models H\varphi_1 \wedge H\varphi_2 \wedge B\neg(\varphi_1 \wedge \varphi_2) \Rightarrow H(\varphi_1 \vee \varphi_2)$
2. $\mathbb{F} \models \neg H\varphi_1 \wedge \neg H\varphi_2 \wedge B\neg(\varphi_1 \wedge \varphi_2) \Rightarrow \neg H(\varphi_1 \vee \varphi_2)$

The first part of Proposition 6 says if both φ_1 and φ_2 are wished, then $\varphi_1 \vee \varphi_2$ should be wished provided that the two wishes contradict each other to the agent. The condition $B\neg(\varphi_1 \wedge \varphi_2)$ is essential, and we can not get $\mathbb{F} \models H\varphi_1 \wedge H\varphi_2 \Rightarrow H(\varphi_1 \vee \varphi_2)$. This is important because, as we demonstrated earlier, it may cause a problem if we accept $\models H\varphi_1 \wedge H\varphi_2 \Rightarrow H(\varphi_1 \vee \varphi_2)$ generally.

Let us consider the second part of Proposition 6. It seems that if both φ_1 and φ_2 are not good news to Alice, then $\varphi_1 \vee \varphi_2$ should not be good news to Alice. However, by Proposition 6, from the condition that both φ_1 and φ_2 are not good news to Alice, we can not conclude $\varphi_1 \vee \varphi_2$ is not good news unless we have an extra condition such as Alice believes that φ_1 and φ_2 contradict each other, i. e ., $B\neg(\varphi_1 \wedge \varphi_2)$. We argue that the condition $B\neg(\varphi_1 \wedge \varphi_2)$ is necessary; there is indeed some case that $\varphi_1 \vee \varphi_2$ is good news to Alice but neither φ_1 nor φ_2 is. Suppose that Alice is a vegetarian and has been invited to a dinner. The invitation letter indicates that exactly one kind of food from beef, chicken, pumpkin, and tomato will be served and each of them will be chosen with the same probability. Let b, c, p and t represent that beef, chicken, pumpkin, and tomato are chosen, respectively. Assume that Alice likes pumpkin and tomato to the same extent; so she dislikes beef and chicken. In this scenario, we can reasonably conclude the following claims.

- $\neg H(b \vee p)$: It is not good news to Alice that beef or pumpkin will be served.

  The reason for the above claim is that after Alice is told such news she can not expect to get her favored food with more probability. The probability that she gets her favored food is still $1/2$, the same as that without being told ($b \vee p$).

- $\neg H(b \vee t)$: It is not good news to Alice that beef or tomato will be served.

  The argument for the reasonability of this claim is the same as for the claim in the first item.

- $H((b \vee p) \vee (b \vee t))$: It is good news to Alice that beef or pumpkin or tomato will be served.

  If Alice is told $(b \vee p) \vee (b \vee t)$, she knows that beef or pumpkin or tomato will be served, but chicken will not; as a result, she can conclude that the probability that she gets her favored food becomes $2/3$ instead of $1/2$. Therefore, we can reasonably conclude that $H((b \vee p) \vee (b \vee t))$, which means that Alice would be happier being told $(b \vee p) \vee (b \vee t)$.

It is thus reasonable for an agent to desire $\varphi \vee \psi$ even though the agent neither desires φ nor ψ; indeed, the intuition is not always reliable.

Corollary 7 *For any linear and introspective frame $\mathbb{F}$, we have that*

1. $\mathbb{F} \models H\varphi_1 \wedge \neg H\varphi_2 \wedge B(\varphi_2 \Rightarrow \varphi_1) \Rightarrow H(\varphi_1 \wedge \neg\varphi_2)$
2. $\mathbb{F} \models \neg H\varphi_1 \wedge H\varphi_2 \wedge B(\varphi_2 \Rightarrow \varphi_1) \Rightarrow \neg H(\varphi_1 \wedge \neg\varphi_2)$

Completeness Result

We now give an axiom system or proof theory for beliefs and pro attitudes, denoted by **BPA**. Proof theory **BPA** consists of the following reasoning rules and axiom schemas.

- Propositional tautologies and inference rules.
- K45 for B,
  - $\vdash B(\varphi \Rightarrow \psi) \wedge B\varphi \Rightarrow B\psi$
  - $\vdash (B\varphi \Rightarrow BB\varphi) \wedge (\neg B\varphi \Rightarrow B\neg B\varphi)$

- Unit-zero-exclusion for H: $\vdash \neg H true \wedge \neg H\neg true$
- BH-introspection,
  - $\vdash (H\varphi \Rightarrow BH\varphi) \wedge (\neg H\varphi \Rightarrow B\neg H\varphi)$
  - $\vdash B(\varphi \Leftrightarrow \psi) \Rightarrow (H\varphi \Leftrightarrow H\psi)$
- $\vdash H\varphi_1 \wedge H\varphi_2 \wedge B\neg(\varphi_1 \wedge \varphi_2) \Rightarrow H(\varphi_1 \vee \varphi_2)$
- $\vdash \neg H\varphi_1 \wedge \neg H\varphi_2 \wedge B\neg(\varphi_1 \wedge \varphi_2) \Rightarrow \neg H(\varphi_1 \vee \varphi_2)$
- Inference rules for B: $\frac{\vdash \varphi}{\vdash B\varphi}$

Theorem 8 *Proof theory **BPA** is sound and complete with respect to linear, introspective, and unit-zero-exclusive models.*

The soundness part is straightforward. The completeness is a little more complex, and we only present a succinct outline of its proof. As for completeness, we will show that every consistent set of formulas can be satisfied in some linear, introspective, and unit-zero-exclusive model. By the strategies for constructing "canonical models" in both normal and non-normal modal logics (Chellas 1980), we build our "canonical model" $\mathbb{M}_c = \langle W_c, R_c, N_c, V_c \rangle$ as follows:

1. Let W_c be the collection of all maximal consistent formula sets.

2. For every formula set Γ, let $\Gamma/B = \{\varphi \mid B\varphi \in \Gamma\}$, and W_c^Γ the collection of those maximal consistent sets that contain Γ/B.

3. Let R_c be a relation on W_c such that, for all $w, w' \in W_c$, wR_cw' iff $w' \in W_c^w$.

4. For every $p \in \mathbf{At}$, $V_c(p) = \{w \mid w \in W_c \text{ and } p \in w\}$.

5. For every formula φ, let
   $$|\varphi| = \{w \mid w \in W_c \text{ and } \varphi \in w\}.$$
   It follows that $|\varphi_1| = |\varphi_2|$ implies that, for every $w \in W_c$, $H\varphi_1 \in w$ iff $H\varphi_2 \in w$.

6. Neighborhood Function N_c will be constructed such that, for each $w \in W_c$ and each formula φ, $|\varphi| \in N_c(w)$ iff $H\varphi \in w$.

The following claims play an important role in the completeness proof:

1. For every $w \in W_c$ and every φ, if $\neg B\varphi \in w$, then there is a $w' \in W_c^w$ such that $\neg\varphi \in w'$.

2. For every $w \in W_c$, every φ_1 and every φ_2, if $|\varphi_1| \cap W_c^w \subseteq |\varphi_2| \cap W_c^w$ then $\varphi_1 \Rightarrow \varphi_2 \in w/B$.

3. For every φ_1 and every φ_2, $|\varphi_1| = |\varphi_2|$ implies that, for every $w \in W_c$, $H\varphi_1 \in w$ iff $H\varphi_2 \in w$.

Now we complete the details of the construction $N_c(w)$ for every $w \in W_c$. For every $X \subseteq W_c$, if X is of the form $|\varphi|$, then $X \in N_c(w)$ iff $H\varphi \in w$. Let $\mathcal{U}$ be the collection of those subsets of W_c^w that are not of the form $|\varphi| \cap W_c^w$. By the choice axiom, we can suppose that there is a well-order, say $<_\mathcal{U}$, on $\mathcal{U}$. For all $X \in \mathcal{U}$, we will determine whether X is in $N_c(w)$ by induction on $<_\mathcal{U}$.

We first introduce auxiliary notations $cl^+(\mathcal{W}, \mathcal{W}')$ and $cl^-(\mathcal{W}, \mathcal{W}')$. For a pair of collections $\mathcal{W}, \mathcal{W}' \subseteq W_c^w$, let $cl^+(\mathcal{W}, \mathcal{W}')$ and $cl^-(\mathcal{W}, \mathcal{W}')$ be the two smallest sets such that

1. $\mathcal{W} \subseteq cl^+(\mathcal{W}, \mathcal{W}')$ and $\mathcal{W}' \subseteq cl^+(\mathcal{W}, \mathcal{W}')$
2. For $X_1, X_2 \in cl^+(\mathcal{W}, \mathcal{W}')$, if $X_1 \cap X_2 = \{\}$, then $X_1 \cup X_2 \in cl^+(\mathcal{W}, \mathcal{W}')$.
3. For $X_1, X_2 \in cl^-(\mathcal{W}, \mathcal{W}')$, if $X_1 \cap X_2 = \{\}$, then $X_1 \cup X_2 \in cl^-(\mathcal{W}, \mathcal{W}')$.
4. For $X_1 \in cl^+(\mathcal{W}, \mathcal{W}')$ and $X_2 \in cl^-(\mathcal{W}, \mathcal{W}')$, if $X_2 \subseteq X_1$, then $X_1 - X_2 \in cl^+(\mathcal{W}, \mathcal{W}')$.
5. For $X_1 \in cl^+(\mathcal{W}, \mathcal{W}')$ and $X_2 \in cl^-(\mathcal{W}, \mathcal{W}')$, if $X_1 \subseteq X_2$, then $X_2 - X_1 \in cl^-(\mathcal{W}, \mathcal{W}')$.

For every $X \in \mathcal{U}$ we define two collections of subsets W_c, denoted by $\mathcal{W}_X$ and $\mathcal{W}'_X$, as follows:

1. If X is the least $(<_{\mathcal{U}})$ among $\mathcal{U}$, let

$$\mathcal{W}_0 = \{|\varphi| \cap W_c^w \mid H\varphi \in w\}$$
$$\mathcal{W}'_0 = \{|\varphi| \cap W_c^w \mid H\varphi \notin w\}$$
$$\mathcal{W}_X = cl^+(\mathcal{W}_0, \mathcal{W}'_0 \cup \{X\})$$
$$\mathcal{W}'_X = cl^-(\mathcal{W}_0, \mathcal{W}'_0 \cup \{X\})$$

2. If X is not the least $(<_{\mathcal{U}})$ among $\mathcal{U}$, there are two cases:
   (a) If there is $Y <_{\mathcal{U}} X$ such that $X \in \mathcal{W}_Y$ or $X \in \mathcal{W}'_Y$, then $\mathcal{W}_X = \bigcup_{Y <_{\mathcal{U}} X} \mathcal{W}_Y$ and $\mathcal{W}'_X = \bigcup_{Y <_{\mathcal{U}} X} \mathcal{W}'_Y$.
   (b) Otherwise, let
   $$\mathcal{W}_X = cl^+(\bigcup_{Y <_{\mathcal{U}} X} \mathcal{W}_Y, \bigcup_{Y <_{\mathcal{U}} X} \mathcal{W}'_Y \cup \{X\})$$
   $$\mathcal{W}'_X = cl^-(\bigcup_{Y <_{\mathcal{U}} X} \mathcal{W}_Y, \bigcup_{Y <_{\mathcal{U}} X} \mathcal{W}'_Y \cup \{X\})$$

Now we can completely identify the neighborhood function N_c as follows. For every $w \in W_c$, and every $X \subseteq W_c$,

- if X is of the form $|\varphi|$, then $X \in N_c(w)$ iff $H\varphi \in w$;
- otherwise, we have that $Y = X \cap W_c^w \in \mathcal{U}$ and let $X \in N_c(w)$ iff $Y \in \mathcal{W}_Y$.

The Completeness Proof: We first show that model $\mathbb{M}_c$ is linear, introspective, and unit-zero-exclusive, and for any formula φ, $(\varphi)^{\mathbb{M}_c} = |\varphi|$. Given an arbitrary consistent set Γ_0 of formulas in $\mathcal{L}^{B,H}(\mathbf{At})$, we can extend Γ_0 to a maximal consistent formula set w_0. Because for any formula φ, $(\varphi)^{\mathbb{M}_c} = |\varphi|$, we have that $\mathbb{M}_c, w_0 \models w_0$, and hence $\mathbb{M}_c, w_0 \models \Gamma_0$. Thus, Γ_0 is satisfiable. This completes the proof. ∎

Generating the Models by Probabilistic Programming with Utilities

Computational grounding is an important requirement for logical theories of agency. To show our logic is computationally grounded, we now give our logic an interpretation in terms of some concrete computational model.

We consider those probabilistic programs π. The semantics of π is characterized by a set W_π of possible runs as well as probabilistic distributive function Pr_π over W_π. Suppose that from program π, we can derive a utility function U_π such that for each run $r \in W_\pi$, $U_\pi(r)$ is a real number representing the reward for the agent finishing run r. Also, we assume the agent has incomplete information about the environment so that for every run r, there are some runs that are indistinguishable to the agent. We use $r \approx r'$ to denote that runs r and r' are indistinguishable to the agent, and $[r]_\approx$ the set of all runs indistinguishable from r. For simplicity, we assume that W_π is a finite set, i.e., there are only finitely many possible runs for the program.

We derive the corresponding hybrid neighborhood frame $\mathbb{M}_\pi = (W_\pi, R_\pi, N_\pi)$ as follows.

- W_π is the set of all possible runs.
- $rR_\pi r'$ iff $r \approx r'$ and $Pr_\pi(r') > 0$.
  Thus, the agent who is currently at run r, believes a run r' is possible iff r' is indistinguishable from r and the probability of the run is non-zero.
- Given $r \in W_\pi$ and $X \subseteq W_\pi$, we have that $X \in N_\pi(r)$ iff the following holds

$$\frac{\sum_{r' \in X, \, r' \approx r} Pr_\pi(r')U_\pi(r')}{\sum_{r' \in X, \, r' \approx r} Pr_\pi(r')} > \frac{\sum_{r' \approx r} Pr_\pi(r')U_\pi(r')}{\sum_{r' \approx r} Pr_\pi(r')}.$$

  Notice that $\frac{\sum_{r' \approx r} Pr_\pi(r')U_\pi(r')}{\sum_{r' \approx r} Pr_\pi(r')}$ is the interests the agent expects to get because the agent believes only those runs in $[r]_\approx$ are possible, while $\frac{\sum_{r' \in X, \, r' \approx r} Pr_\pi(r')U_\pi(r')}{\sum_{r' \in X, \, r' \approx r} Pr_\pi(r')}$ is the interests the agent expects to get when being told only runs in X are possible. Therefore, the intuitive meaning of $X \in N_\pi(r)$ is that the agent gains the expected interests when being told only runs in X are possible.

By the construction above, we have the following claim.

Proposition 9 $\mathbb{M}_\pi$ *is a linear, introspective, and unit-zero-exclusive model.*

Thus, we have established a relationship between the present semantic model for beliefs and pro attitudes and a concrete computational model. We argue that if a semantic model of agency does not have such a relationship to any concrete computational model, then the semantic model is just a purely artificial one and can not be thought of as natural and rational. For example, an alternative semantic model can be artificially constructed for some logic theory of beliefs and pro attitudes with an intuitively valid (but problematic to some extent!) property: $\vdash H\varphi_1 \wedge H\varphi_2 \Rightarrow H(\varphi_1 \vee \varphi_2)$. The related soundness and completeness results can also be carried out. However, this purely artificial semantic model does not guarantee the rationality of the alternative logic, because there might not be a relationship between the semantic model and any concrete computational one.

Related Work

In this section, we discuss some related work from decision theory and agent theory.

Preferences and Utilities Several logics of desires and goals have been proposed for modeling preferences and utilities (Doyle, Shoham, & Wellman 1991; Lang, van der Torre, & Weydert 2002; Boutilier 1994; Broersen, Dastani, & van der Torre 2002). These logics are not modal logics and their focuses are on representational issues, offering AI languages to represent preferences and utilities compactly and implicitly.

Modal logics of Beliefs, Desires and Intentions Two influential modal logics of beliefs and pro attitudes are the theory of intention (Cohen & Levesque 1990) and the formalism of the belief-desire-intention paradigm (Rao & Georgeff

1998). (Cohen & Levesque 1990) did not provide a complete axiom system, while BDI_CTL (Rao & Georgeff 1998) can be reduced to standard modal logics such as mu-calculus (Schild 2000) and thus can be axiomatizable. However, both of them suffer from the so-called side-effect problem (to some extent). Moreover, as argued by (Wooldridge 2000; van der Hoek & Wooldridge 2003), they are not computationally grounded.

Representationalism (Konolige & Pollack 1993) gave a representationalist theory of intention. They did make their modal logic of intention fully side-effect free. Nevertheless, they did not provide a complete axiom system and did not consider the requirement of "computational grounding".

Computational Grounding Besides epistemic logics, another computationally grounded logic is $\mathcal{VSK}$ logic (Wooldridge & Lomuscio 2001), which enables us to represent what is *visible* of the environment to individual agents, what these agents actually *perceive* (see), and what the agents actually *know* about the environment. However, this work does not concern an agent's pro attitudes. (Su *et al.* 2005; 2006) presented a computationally grounded model of knowledge, belief, desire and intention, called the interpreted KBDI-system model. Unfortunately, the notions of desire and intention they characterized are based on normal modal logics and thus suffer from the side-effect problem.

Conclusion

In this paper, we have presented a new modal logic formalizing agents' pro attitudes, based on neighborhood models. The distinguishing features of this logic are on three aspects:

- Firstly, this logic naturally satisfies *Bratman's requirements* for agents' beliefs and pro attitudes. We demonstrate, by a realistic example, some intuitive properties may surprisingly cause a problem, which, to the best of our knowledge, have not been discussed before (see the discussions below Proposition 6).

- Secondly, we have obtained a sound and complete axiom system for capturing valid properties of beliefs and pro attitudes. Moreover, based on the additivity (or linearity) of the underlying probability and utility expectation functions, we introduce the notion of *linear neighborhood frame* for obtaining the semantic model for beliefs and pro attitudes, which was not investigated in the literature of modal. We are not just to borrow from modal logics, but bring a new member to the family of non-normal modal logics.

- Finally, the semantic models of this logic can be naturally derived from probabilistic programming with utilities and this logic can be thought of as computationally grounded, which takes a promising step towards the challenging open issue of giving a computationally grounded semantics to goals (Wooldridge 2000).

As for future work, we will consider some specific classes of pro attitudes such desires, goals and intentions, and relate them to agents' actions. The ongoing work is to add the time dimension to the present semantic framework and to explore the model checking problem.

References

Boutilier, C. 1994. Towards a logic of qualitative decision theory. In *Proc. of KR'94*, 75–86.

Bratman, M. 1987. *Intention, Plans, and Practical Reason*. Cambridge, MA, USA: Harvard University Press.

Broersen, J.; Dastani, M.; and van der Torre, L. W. N. 2002. Realistic desires. *Journal of Applied Non-Classical Logics* 12(2).

Chellas, B. F. 1980. *Modal Logic*. Cambridge University Press.

Cohen, P., and Levesque, H. 1990. Intension is choice with commitment. *Artificial Intelligence* 42:23–261.

Doyle, J.; Shoham, Y.; and Wellman, M. 1991. The logic of relative desires. In *Sixth International Symposium on Methodologies for Intelligent Systems*.

Fagin, R.; Halpern, J.; Moses, Y.; and Vardi, M. 1995. *Reasoning about knowledge*. Cambridge, MA: MIT Press.

Halpern, J., and Vardi, M. 1986. The complexity of reasoning about knowledge and time: extended abstract. In *Proc. STOC-86*.

Halpern, J., and Zuck, L. 1992. A little knowledge goes a long way: Simple knowledge based derivations and correctness proofs for a family of protocols. *Journal of the ACM* 39(3):449–478.

Konolige, K., and Pollack, M. E. 1993. A representationalist theory of intention. In *IJCAI-93*, 390–395.

Lang, J.; van der Torre, L.; and Weydert, E. 2002. Utilitarian desires. *Autonomous Agents and Multi Agent systems* 5(3):329–363.

Montague, R. 1970. Universal grammar. *Theoria* 36:373–398.

Rao, A., and Georgeff, M. 1998. Decision procedures for BDI logics. *Journal of Logic and Computation* 8(3):293–344.

Schild, K. 2000. On the relationship between BDI logics and standard logics of concurrency. *Autonomous Agents and Multi-Agent Systems* 3:259–283.

Scott, D. Advice in modal logic. In Lambert, K., ed., *Philosophical Problems in Logic*.

Su, K.; Sattar, A.; Wang, K.; Luo, X.; Governatori, G.; and Padmanabhan, V. 2005. The observation-based model for BDI-agents. In *AAAI-05*, 190–195.

Su, K.; Yue, W.; Sattar, A.; Orgun, M. A.; and Luo, X. 2006. Observation-based logic of knowledge, belief, desire and intention. In *Proc. of KSEM-06*, 366–378.

van der Hoek, W., and Wooldridge, M. 2003. Towards a logic of rational agency. *L.J. of IGPL* 11(2):135–159.

Wooldridge, M., and Lomuscio, A. 2001. A computationally grounded logic of visibility, perception, and knowledge. *Logic Journal of the IGPL* 9(2):273–288.

Wooldridge, M. 2000. Computationally grounded theories of agency. In Durfee, E., ed., *ICMAS-00*, 13–22. IEEE Press.