The Interpreted System Model of Knowledge, Belief, Desire and Intention

Kaile Su1, Xiangyu Luo1, Abdul Sattar1, Mehmet A Organ2
1Institute for Integrated and Intelligent Systems, Griffith University, Brisbane, Australia
2Department of Computing, Macquarie University, Sydney, Australia
{k.su, x.luo, a.sattar}@griffith.edu.au; mehmet@comp.mq.edu.au

ABSTRACT
We present a new model of knowledge, belief, desire and intention, called the interpreted KBDI-system model (or KBDI-model for short). The key point of the interpreted KBDI-system model is that we express an agent’s knowledge, belief, desire and intention as a set of runs (computing paths), which is exactly a system in the interpreted system model, a well-known agent model due to Halpern and his colleagues. Our KBDI-model is computationally grounded in that we are able to associate a KBDI-model with a computer program, and formulas involving agents’ knowledge, belief, desire (goal) and intention, can be understood as properties of program computations. We present a sound and complete proof system with respect to our KBDI-model and explore how symbolic model checking techniques can be applied to model checking multi-agent systems with KBDI-models.

1. INTRODUCTION
The possible worlds semantics [1] is a fruitful approach to formalizing agent systems via modal logics. The well-known theory of intention [3] and the formalism of the belief-desire-intention paradigm [10], for example, are along this line. However, it is still not very clear how to obtain concrete agent models with the belief, desire and intention accessibility relations from specific agent programs.

The interpreted system model [4] offers a natural interpretation, in terms of the states of computer processes, to S5 epistemic logic. The salient point of the interpreted system model is that we are able to associate an interpreted system with a computer program, and formulas in epistemic logic that are valid with respect to the interpreted system model. In this sense, the interpreted system model is computationally grounded [14].

The aim of this paper is to present a computationally grounded model of knowledge, belief, desire and intention, called the inter-

*An alternative version of a paper in Proceedings of AAMAS 2006. This work was supported by the Australian Research Council grant DP0452628, and partially by the NSFC grants 60496327, 10410638 and 60473004.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 200X ACM X-XXXX-XX-X/XX/XX ...$5.00.

We present a new model of knowledge, belief, desire and intention, called the interpreted KBDI-system model (or KBDI-model for short). The key point of the interpreted KBDI-system model is that we express an agent’s knowledge, belief, desire and intention as a set of runs (computing paths), which is exactly a system in the interpreted system model, a well-known agent model due to Halpern and his colleagues. Our KBDI-model is computationally grounded in that we are able to associate a KBDI-model with a computer program, and formulas involving agents’ knowledge, belief, desire (goal) and intention, can be understood as properties of program computations. We present a sound and complete proof system with respect to our KBDI-model and explore how symbolic model checking techniques can be applied to model checking multi-agent systems with KBDI-models.

1. INTRODUCTION
The possible worlds semantics [1] is a fruitful approach to formalizing agent systems via modal logics. The well-known theory of intention [3] and the formalism of the belief-desire-intention paradigm [10], for example, are along this line. However, it is still not very clear how to obtain concrete agent models with the belief, desire and intention accessibility relations from specific agent programs.

The interpreted system model [4] offers a natural interpretation, in terms of the states of computer processes, to S5 epistemic logic. The salient point of the interpreted system model is that we are able to associate an interpreted system with a computer program, and formulas in epistemic logic that are valid with respect to the interpreted system model. In this sense, the interpreted system model is computationally grounded [14].

The aim of this paper is to present a computationally grounded model of knowledge, belief, desire and intention, called the inter-
sents one possible computation of a system and a system may have a number of possible runs, so we say a system is a set of runs. Assume that we have a set $Φ$ of primitive propositions, which we can think of as describing basic facts about the system. An interpreted system $I$ consists of a pair $(R, π)$, where $R$ is a set of runs over a set of global states and $π$ is a valuation function, which gives the set of primitive propositions true at each point in $R$.

For every agent $i$, let the notation $(r, u) ∼_i (r', u')$ denote that $r_i(u) = r'_i(u)$. Intuitively, $(r, u) ∼_i (r', v')$ means that $(r, u)$ and $(r', v')$ are indistinguishable to agent $i$. We also use the notation $(r, u) ∼^spr_i (r', v')$ to denote that $u = v$ and, for every $j ≤ u$, $r_i(j) = r'_i(j)$ (here $spr$ stands for synchronous systems with perfect recall).

Let $K_{L_o}$ denote the language of propositional logic augmented by the future-time connectives $\Box$ (next) and $\Diamond$ (until) and a modal (knowledge) operator $K_i$ for each agent $i$. The language $K_{L_o}$ can be interpreted by using an interpreted system. The related satisfaction relation $|=_{K_{L_o}}$ is as follows: Given $I = (R, π)$ and a point $(r, u)$ in $R$, we define $(I, r, u) |=_{K_{L_o}} ϕ$ by induction on the structure of $ϕ$. When $ϕ$ is of the form $K_i ϕ_i$, $(I, r, u) |=_{K_{L_o}} ϕ$ iff $(I, r', v) |=_{K_{L_o}} ϕ$ for all $(r', v)$ such that $(r, u) ∼_i (r', v)$.

### 2.2 Interpreted KBDI-Systems

Given a set $G$ of global states and system $K$ over $G$, an agent’s mental state over system $K$ is a tuple $(B, D, I)$, where $B$, $D$, and $I$ are sets (sets of runs over $G$) such that $I ⊆ K$ and $D ⊆ B ⊆ K$. A KBDI-system is a structure $(K, M_1, . . . , M_n)$, where $K$ is a system and for every $i$, $M_i$ is agent $i$’s mental state over $K$.

Assume that we have a set $Φ$ of primitive propositions which describe basic facts about agents and their environment. An interpreted KBDI-system $I$ consists of a pair $(S, π)$, where $S$ is a KBDI-system and $π$ is a valuation function, which gives the set of primitive propositions true at each point in $G$.

### 2.3 Syntax

Given a set $Φ$ of propositional atoms, the language of KBDI logic is defined by the following BNF notations:

\[
\text{wff} ::= \text{any element of } Φ | \neg (\text{wff}) | (\text{wff}) \land (\text{wff}) | (\text{wff}) \lor (\text{wff}) | (\text{wff}) \rightarrow (\text{wff}) | (\text{wff}) \leftrightarrow (\text{wff}) | K_i (\text{wff}) | B_i (\text{wff}) | D_i (\text{wff}) | I_i (\text{wff})
\]

Informally, $K_i ϕ$, $B_i ϕ$, and $D_i ϕ$ means that agent $i$ knows, believes and desires $ϕ$, respectively, while $I_i ϕ$ denotes that $ϕ$ holds under the assumption that agent $i$ acts based on his intention. The formulas not containing modalities $K_i$, $B_i$, $D_i$, and $I_i$ are called linear-temporal logic (LTL) formulas.

### 2.4 Semantics

We now proceed to interpret KBDI logic formulas in terms of interpreted KBDI-systems. In the following, we inductively define the satisfaction relation $|=_O$ between a formula $ϕ$ and a pair of interpreted KBDI-system and a point. Given an interpreted KBDI-system $I = (S, π)$, suppose that $S = (K, M_1, . . . , M_n)$ and for every $i$, $M_i = (B_i, D_i, I_i)$. Let $r$ be a run in $K$ and $u$ a natural number, then we have that:

- $(I, r, u) |=_O K_i ϕ$ iff $(I, r', v) |= O ϕ$ for those $(r', v)$ such that $r' ∈ K$ and $(r, u) ∼_i (r', v)$;
- $(I, r, u) |=_O B_i ϕ$ iff $(I, r', v) |= O ϕ$ for those $(r', v)$ such that $r' ∈ B_i$ and $(r, u) ∼_i (r', v)$;
- $(I, r, u) |=_O D_i ϕ$ iff $(I, r', v) |= O ϕ$ for those $(r', v)$ such that $r' ∈ D_i$ and $(r, u) ∼_i (r', v)$;
- $(I, r, u) |=_O I_i ϕ$ iff $(I, r', v) |= O ϕ$ for those $(r', v)$ such that $r' ∈ I_i$ and $(r, u) ∼_i (r', v)$;

The semantics of atomic formulas $ϕ$ or formulas of the form $¬ϕ$, $ϕ \land ϕ'$, $ϕ \lor ϕ'$, or $ϕ \rightarrow ϕ'$ can be dealt with in the usual manner.

We use $|=_O φ$ to denote that $ϕ$ is valid in every interpreted KBDI-system.

According to our definition, $D_i φ$ is true iff $φ$ is true along those runs that are desirable to agent $i$ and consistent with agent $i$’s observations. Thus, $D_i φ$ intuitively means that agent $i$’s goal implies that formula $φ$ holds.

For those agents with perfect recall and a global clock, we may use $¬\sim^spr_i$ instead of $\sim_i$ to interpret those formulas with modalities $B_i$, $D_i$ and $I_i$ and get an alternative satisfaction relationship $|=^{spr}_O$.

**Proposition 1.** The following axioms are valid with respect to both $|=_O$ and $|=^{spr}_O$:

- $K_i φ \Rightarrow φ$
- $Δ_i (φ = ψ) \Rightarrow (Δ_i φ = Δ_i ψ)$
- $Δ_i φ \Rightarrow K_i Δ_i φ = K_i −Δ_i φ$

where $Δ$ stands for $K$, $B$, $D$ or $I$.

- Relationship between knowledge, belief, desire and intention $B_i φ \Rightarrow D_i φ \Rightarrow K_i φ \Rightarrow B_i φ \Rightarrow K_i φ \Rightarrow I_i φ$

- Temporal operators
  - $O(φ = ψ) \Rightarrow (O φ = O ψ)$
  - $O(¬φ) \Rightarrow ¬ O φ$
  - $φ ≤ ψ \Rightarrow ψ ≤ (φ ∧ O(φ ≤ ψ))$

**Proposition 2.** The following axioms are valid with respect to $|=^{spr}_O$: $Δ_i Δ_i φ = Δ_i φ$, where $Δ$ stands for $K$, $B$, $D$ and $I$.

The formula $D_i φ \Rightarrow φ$ holds if and only if $O ϕ$ holds at the next point in time, then at the next point in time her goal will imply $ϕ$, that is, agent $i$ persists on her goal.

### 2.5 Proof system

We now discuss a proof system, called the KBDI proof system, for those agents with perfect recall and a global clock. The proof system contains the axioms of propositional calculus plus those in Propositions 1 and 2. It is closed under the propositional inference rules plus for every agent $i$: $\frac{\text{spr}}{K_i ϕ}$.

**Theorem 3.** The KBDI proof system for agents with perfect recall and a global clock is sound and complete with respect to interpreted KBDI-systems with satisfaction relation $|=^{spr}_O$.

### 3. MODEL CHECKING KBDI LOGIC

The model checking problem is to give an algorithm to test whether a model satisfies a concerned specification expressed by a formula.

However, in order to make our model checking algorithm practically useful, we must consider from where our model, an interpreted KBDI-system, comes. It would be satisfactory if we can derive our model from a program implemented in, say C or Java. However, to simplify the matter, we may consider some abstract programs such as finite-state programs, which are expressive enough from the standpoint of theoretical computer science. Moreover, to make our model checking system practically efficient, we use symbolic model checking techniques. Thus, a finite-state program in our approach is represented in a symbolic way.

### 3.1 Symbolic Representation of Interpreted KBDI-System Model

A system as a set of infinite runs is not well suited to model checking directly as it is generally applied to finite state systems. However, we can represent an interpreted system as a finite-state program $(G, G_0, B, V)$, where $G$ is a set of states, $G_0$ a set of initial states, and $R$ a total “next time” relation, and $V$ associates
each state with a truth assignment function. A set of infinite runs is then obtained by “unwinding” the relation $R$ starting from initial states in $G_0$.

We can present a symbolic representation of a finite-state program $(G, G_0, V)$ as $(x, \theta(x), \tau(x, x'))$. Also, we represent a system as a (symbolic) finite-state program with $n$ agents, which is defined as a tuple $P = (x, \theta(x), \tau(x, x'), O_1, \ldots, O_n)$, where

1. $x$ is a set of system variables;
2. $\theta$ is a boolean formula over $x$, called the initial condition;
3. $\tau$ is a boolean formula over $x \cup x'$, called the transition relation; and
4. for each $i$, $O_i \subseteq x$, containing agent $i$’s local variables, or observable variables.

For convenience, we may use $P(\theta, \tau)$ to denote a finite-state program with $n$ agents $(x, \theta(x), \tau(x, x'), O_1, \ldots, O_n)$, if $x$ and $O_1, \ldots, O_n$ are clear from the context.

Given a finite-state program $P(\theta, \tau)$ with $n$ agents, we define an agent’s internal program over $P(\theta, \tau)$ as a tuple $(P(\theta_i, \tau_1), P(\theta_2, \tau_2), P(\theta_3, \tau_3))$

where, for $j = 1, 2, 3$, $(\theta_j \Rightarrow \theta) \wedge \tau_j \Rightarrow \tau$ is valid, and $(\theta_2 \Rightarrow \theta_1) \wedge \tau_2 \Rightarrow \tau_1$ is valid.

Clearly, an agent’s internal program is exactly related to an agent’s mental state. Thus, we define a symbolic BDI-program with $n$ agents as a tuple $P_n = (P_1, P_2, \ldots, P_n)$, where $P_1$ is a finite-state program with $n$ agents and for each agent $i$, $P_i$ is agent $i$’s internal program over $P_n$. We use $I_{P_n}$ to denote the corresponding interpreted KBDI-system.

### 3.2 Simplified BDI-Programs

To obtain a BDI-program with $n$-agents, one should define $1 + 3n$ finite-state programs. Our experience with a previous version of MCKBDI indicated that it is inconvenient for a programmer to code a BDI-program with $n$-agents because there may be many finite-state programs involved. Therefore, we should simplify the definition of a BDI-program with $n$-agents.

**Definition 4.** Let $P = (x, \theta(x), \tau(x, x'), O_1, \ldots, O_n)$ be a finite-state program with $n$-agents, where, for each agent $i$, $O_i$ contains at least 3 particular boolean variables $\text{w}_i, \text{w}_D_i$, and $\text{w}_I_i$. A simplified BDI-program with $n$-agents derived from $P$ is a tuple $(P_1, P_2, \ldots, P_n)$, where for each agent $i$, $P_i = (P(\theta_i, \tau_i))$, such that

1. $\theta_i = \theta \wedge \text{w}_i$ and $\tau_i = \tau \wedge \text{w}_i$;
2. $\theta_2 = \theta \wedge \text{w}_2 \wedge \text{w}_D_2$ and $\tau_2 = \tau \wedge \text{w}_2 \wedge \text{w}_D_2$;
3. $\theta_i = \theta \wedge \text{w}_I_i$ and $\tau_i = \tau \wedge \text{w}_I_i$.

It is more convenient for a programmer to code a simplified BDI-program with $n$-agents, and it seems that the notion of a simplified BDI-Program does not lose much generality of that of a BDI-Program.

### 4. CONCLUDING REMARKS

We have proposed a new KBDI-model by using interpreted systems, and developed a computationally grounded BDI logic, called KBDI logic. We interpret BDI logic via two different semantics. One of them is based on the assumption that agents have perfect recall and there is a global clock, with respect to which a sound and complete proof system has been presented. We have explored how symbolic model checking techniques can be applied to model checking BDI-agents and obtained encouraging preliminary experimental results.

This paper is inspired by Su [11] and Su et al [12], which also provide agent models by extending the interpreted system model. However, Su [11] does not consider agents’ belief, desire and intention. The main advantages of the present paper over Su et al [12] are: (1) We introduce, for each agent $i$, not only belief, desire and intention modalities $B_i, D_i$, and $I_i$, but also a knowledge modality $K_i$. Moreover, the resulting axiomatic system is more succinct and proved to be sound and complete. (2) A model checker with KBDI-models, called MCKBDI, is developed [4]. Empirical results are more comprehensive, supported by convincing and motivating examples, where agents’ observations and actions are explicitly defined.

Model checking multi-agent systems has become an active research topic in the community of multi-agent systems. Many efforts have been devoted to model checking knowledge in multi-agent systems [5][13][11][7]. However, comparatively little work has been carried out on model checking BDI-agents. Some general approaches to model checking BDI-agents were proposed in Rao and Georgeff [9] and Benerecetti and Cimatti [11], but no method was given for generating models from actual systems, and so the techniques given there could not easily be applied to verifying real multi-agent systems. In Bordini et al [2], model checking techniques for AgentSpeak(L) [8] have been reported; however, AgentSpeak(L) has very simple semantics. The salient point of our work is that we present a general form of a BDI agent program, from which BDI agent models are generated and specifications in full BDI logics can be verified by symbolic model checking techniques.

### 5. REFERENCES


