Tracing Werewolf game by using extended BDI model

NIDE, Naoyuki
Faculty, Division of Human Life and Environmental Sciences
Nara Women's University
Kita-Uoya Nishimachi, Nara, Japan
E-mail: nide@ics.nara-wu.ac.jp

Shiro Takata
Faculty of Science and Engineering
Kindai University
Kowakae 3–4–1, Higashi-Osaka, Japan

***Abstract*—The werewolf game is a kind of role-playing game in which players have to guess other players' roles from their speech acts (what they say). In this game, players have to estimate other players' beliefs and intentions, and try to modify others' intentions. The BDI model is a suitable model for this game, because it explicitly has notions of mental states, i.e. beliefs, desires and intentions. On the other hand, in this game, players' beliefs are not completely known. Consequently, in many cases it is difficult for players to choose a unique strategy; in other words, players frequently have to maintain probabilistic intentions. However, the conventional BDI model does not have the notion of probabilistic mental states. In this paper, we propose an extension of BDI logic that can handle probabilistic mental states and use it to model some situations in the Werewolf game. We also show examples of deductions concerning that situations. We believe that this study will serve as a basis for developing a Werewolf game agent based on BDI logic.**

I. Introduction

AIs for match-type games such as shogi and go have evolved remarkably. These games are complete information games in that each player ideally knows all the information about the current phase of the game. In contrast, in some games, players can only get incomplete information about the game. Such games are called incomplete information games, and the Werewolf game (originally also known as Mafia) is one such game.

Some attempts have already been made to develop agents for playing Werewolf, and they typically make use of machine learning. Yet, we consider that the BDI model is more useful for developing such agents. In the Werewolf game, due to the incompleteness of information, players have to estimate other players' beliefs and intentions, and try to modify others' intentions. The BDI model is useful for modeling such circumstances, since it explicitly has notions of mental states such as beliefs, desires and intentions.

However, the conventional BDI model does not have the notion of probabilistic mental states, especially probabilistic intentions. In other words, it is considered normal that a BDI agent would choose single plan for its goal at a time and form an intention to commit to the plan. Nonetheless, in Werewolf, due to the incompleteness of information, it is often the case that a player has to choose from a number of strategies. For example, when a player has probabilistic multiple beliefs about the identity of the wolf, then (s)he may have probabilistic multiple intentions as to voting for someone, and may not be able to choose one of them until just before they make the determination. To handle such a situation, it would convenient if the BDI model had notions of probabilistic mental states.

Osawa et al.[1] devised a BDI logic with probabilistic mental states, and used it to model the Werewolf game. They showed reasoning about the players' beliefs. However, they did not give formal semantics or a deduction system. Moreover, their example did not use probabilistic mental states.

We proposed another extended BDI logic with probabilistic mental states, gave its formal semantics and deduction system, and showed several deductions about Werewolf games that used probabilistic mental states[2, in Japanese]. However, it is important to capture the process of state transitions to model the Werewolf game, since it reflects the thinking processes of players. In particular, one agent's beliefs should be able to be modified by another agent's speech acts, i.e. be affected by what the other agent says. Yet, our study mentioned above did not include any examples of this.

In this paper, we show examples of deductions in our logic system, including representations of agents' speech acts and state transitions caused by them. Our logic system, called $\mathcal{TOMATOes}$-$\mathcal{P}$, is an extension of $\mathcal{TOMATO}$[3], which we previously proposed and is itself an extension of the BDI logic. We aim to develop agents for playing Werewolf games based on reasoning, and we believe that this study to be a first step toward this goal.

II. Extended BDI logic

In this section, we introduce our modal logic system $\mathcal{TOMATOes}$-$\mathcal{P}$ ($\mathcal{TOMATO}$ extended with Event Selection and Probabilistic mental states). It is an extension of $\mathcal{TOMATO}$[3] (Theoretical Observation of Multi-Agent system with Tense and Odds) obtained by adding probabilistic mental state operators and event selectability operator.

For simplicity, in regard to mental state consistencies[4], which are considered to be important properties of autonomous agents modeled in the BDI model, we only consider part of them in this paper. To be precise, we do not consider "realism" or any properties derived from it, such as "asymmetry thesis"; e.g. "an agent who does not believe a goal to be eventually achievable will not intend to eventually achieve that goal". We only take into account the introspection axioms; e.g. "any agent has complete beliefs about its mental states"[1].

[1] $\mathcal{TOMATO}$ considers neither realism nor introspection axioms. While our previous work $\mathcal{TOMATOes}$[5] considers both of them, it does not consider probabilistic desires or intentions.

978-1-5090-3931-9/16 $31.00 © 2016 IEEE
DOI 10.1109/ICA.2016.28

A. Syntax

Here, we define the formulas in $\mathcal{TOMMAITOes}$-$\mathcal{P}$. Hereinafter, the word "formula" means one of $\mathcal{TOMMAITOes}$-$\mathcal{P}$ unless expressly stated otherwise. Symbols like x and y are used as the usual variable symbols in first-order predicate logic, while symbols such as $\mathfrak{X}$ and $\mathfrak{Y}$ are variable symbols that express formulas. We call the latter "formula variable symbols". They are mainly used with fixed-point operators.

We choose and fix a first-order language $\mathcal{L}$, an infinite set of formula variable symbols $\mathcal{V}$, a finite set of event constant symbols $\mathcal{E}$ and a finite set of agent constant symbols $\mathcal{A}$. Hereafter, we write $\{p \mid 0 \leqq p \leqq 1\}$ as $[0, 1]$.

We define formulas as follows.

- Any atomic formula in $\mathcal{L}$ is a formula (in $\mathcal{TOMMAITOes}$-$\mathcal{P}$).
- If ϕ, ψ are formulas and x is a variable symbol in $\mathcal{L}$, then $\phi \vee \psi$, $\neg\phi$ and $\forall x\phi$ are formulas.
- If $e \in \mathcal{E}$, then $\mathsf{pos}(e)$ is a formula.
- If Op is one of X^e (where $e \in \mathcal{E}$), BEL^a, DESIRE^a and INTEND^a (where $a \in \mathcal{A}$), and n is a positive integer, and for $i = 1, 2, \cdots, n$, ϕ_i is a formula, $p_i \in [0, 1]$, $r_i \in \{\geqq, >\}$, then $\mathsf{Op}(_{r_1 p_1}\, \phi_1 \,|\, \cdots \,|_{r_n p_n}\, \phi_n)$ is a formula. In particular, if $n = 1$, we omit the outermost parentheses. We call the p_i in this item probability parameters.
- If $\mathfrak{X} \in \mathcal{V}$, then $\mathfrak{X}$ is a formula.
- If ϕ is a formula, $\mathfrak{X} \in \mathcal{V}$, and $\mathfrak{X}$ does not occur negatively in ϕ, then $\mu\mathfrak{X}.\phi$ is a formula. Here, "occur negatively" means that there are odd numbers of $\neg$'s in the path from the root of the tree structure of ϕ to the occurrence of that $\mathfrak{X}$. μ is the least fixed-point operator.

In addition, we introduce operators such as $\wedge$, $\supset$, $\Leftrightarrow$ and $\exists$ as abbreviations in the usual manner, and set standard priority order among operators (e.g. unary operators combine first, $\wedge$ combines before $\vee$, $\supset$ is right associative, and parentheses changes priorities).

For example, $\mathsf{BEL}^a(_{\geqq 0.3}\, \phi_1 \mid _{\geqq 0.5}\, \phi_2)$ is a formula. Intuitively, this formula means "agent a believes ϕ_1 with probability of at least 0.3, and as another possibility, believes ϕ_2 with probability of at least 0.5". DESIRE^a and INTEND^a are similarly interpreted. X^e is an extension of the next-time operator AX in CTL (computation tree logic) with an event e and transition probabilities; for example, $\mathsf{X}^e_{\geqq .3}\, \phi_1$ intuitively means that if an event e occurs, then at the next time point, ϕ_1 holds with probability of at least 0.3. In addition, $\mathsf{pos}(e)$ intuitively means "event e can be executed now".

$\mathsf{BEL}^a_{<p}\, \phi$, $\mathsf{BEL}^a_{=p}\, \phi$ etc. are treated as abbreviations of $\mathsf{BEL}^a_{\geqq 1-p}\, \neg\phi$ and $\mathsf{BEL}^a_{\geqq p}\, \phi \wedge \neg\, \mathsf{BEL}^a_{>p}\, \phi$, etc. The same goes for X^e, etc.

Additionally, we abbreviate $\mathsf{BEL}^a_{\geqq 1}\, \phi$ as $\mathsf{BEL}^a\, \phi$, which denotes a usual (i.e. without probability) belief of agent a (the same goes for DESIRE^a and INTEND^a). Moreover, we abbreviate $\mathsf{X}^e_{\geqq 1}\, \phi$ as $\mathsf{AX}^e\, \phi$, and $\bigwedge_{e \in \mathcal{E}}(\mathsf{pos}(e) \supset \mathsf{AX}^e\, \phi)$ as $\mathsf{AX}\, \phi$. Accordingly, $\mathsf{AX}\, \phi$ denotes "for any possible event between the current time and the next time point, ϕ holds at the latter", and this corresponds to the AX operator of CTL. Because it has the AX operator and a fixed-point operator μ, $\mathcal{TOMMAITOes}$-$\mathcal{P}$ is more expressive than CTL* as a temporal logic. Hereafter, we will assume that all operators of CTL have already been introduced as abbreviations (e.g. $\mathsf{AG}\, \phi$ is an abbreviation of $\neg\mu\mathfrak{X}.(\neg\phi \vee \neg\, \mathsf{AX}\, \neg\mathfrak{X})$).

B. Semantics

1) BDI structure: First we choose and fix the following:

- A set of possible worlds $W(\neq \emptyset)$
- For each $w \in W$, a set of states $St_w(\neq \emptyset)$
- For each $w \in W$ and each $t \in St_w$, an interpretation $i_{w,t}$ of $\mathcal{L}$. We assume that the domain and the interpretation of terms are the same for all w and t.
- For each $w \in W$ and each $t \in St_w$, an non-empty subset $pos_{w,t}$ of $\mathcal{E}$
- For each $w \in W$ and each $e \in \mathcal{E}$, a function $R^e_w : St^2_w \to [0, 1]$ where $\sum_{t' \in St_w} R^e_w(t, t') = 1$ for any $t \in St_w$
- For each $a \in \mathcal{A}$ and each $t \in \bigcup_{w \in W} St_w$, (hereafter, we write $\{w \mid t \in St_w\}$ as W_t) functions $\mathcal{B}^t_a$, $\mathcal{D}^t_a$, $\mathcal{I}^t_a : W^2_t \to [0, 1]$ which satisfy:
  - ⋆ for each $w \in W_t$, $\sum_{w' \in W_t} \mathcal{B}^t_a(w, w') = 1$, and similar for $\mathcal{D}^t_a$, $\mathcal{I}^t_a$
  - ⋆ for each $w, w' \in W_t$ which satisfy $\mathcal{B}^t_a(w, w') > 0$ and each $w'' \in W_t$, $\mathcal{B}^t_a(w, w'') = \mathcal{B}^t_a(w', w'')$, $\mathcal{D}^t_a(w, w'') = \mathcal{D}^t_a(w', w'')$ and $\mathcal{I}^t_a(w, w'') = \mathcal{I}^t_a(w', w'')$

We call a tuple of the above-mentioned components a BDI structure. Intuitively, a state corresponds to a time point in temporal logic, and a possible world is a time tree of states whose edges are $\{(t, t') \mid$ there is some $e \in pos_{w,t}$ where $R^e_w(t, t') > 0\}$. $pos_{w,t}$ is a set of events which can be currently executed, and $R^e_w(t, t') = p$ means that if an event e is executed at state t, the next time is t' with probability p. $\mathcal{B}^t_a$, $\mathcal{D}^t_a$, and $\mathcal{I}^t_a$ are accessibility relations (with possibilities) between possible worlds, which represent the belief, desire and intention of agent a with possibilities at time t. $\mathcal{B}^t_a(w, w') > 0$ roughly corresponds to the existence of the accessibility relation $w\ \mathcal{B}^t_a\ w'$ in the usual Kripke semantics.

2) Interpretation of formulas: Hereafter, we write $\{(w, t) \mid w \in W,\ t \in St_w\}$ as Swt. Given a BDI structure M and a function $f_{\mathcal{V}} : \mathcal{V} \to 2^{Swt}$, we define the interpretation $[\![\phi]\!]_{\langle M, f_{\mathcal{V}}\rangle}$ of a formula ϕ as follows (note that $[\![\phi]\!]_{\langle M, f_{\mathcal{V}}\rangle} \subseteq Swt$). We say that ϕ holds at a state t of a world w when $[\![\phi]\!]_{\langle M, f_{\mathcal{V}}\rangle} \ni (w, t)$. If $[\![\phi]\!]_{\langle M, f_{\mathcal{V}}\rangle} = Swt$ for any M and $f_{\mathcal{V}}$, we say that ϕ is valid.

- If ϕ is an atomic formula, $[\![\phi]\!]_{\langle M, f_{\mathcal{V}}\rangle} = \{(w, t) \mid \phi$ is true w.r.t. $i_{w,t}\}$
- $[\![\phi \vee \psi]\!]_{\langle M, f_{\mathcal{V}}\rangle} = [\![\phi]\!]_{\langle M, f_{\mathcal{V}}\rangle} \cup [\![\psi]\!]_{\langle M, f_{\mathcal{V}}\rangle}$
- $[\![\neg\phi]\!]_{\langle M, f_{\mathcal{V}}\rangle} = Swt \setminus [\![\phi]\!]_{\langle M, f_{\mathcal{V}}\rangle}$

- $[\![\forall x\phi]\!]_{\langle M,f_{\mathcal{V}}\rangle} = \bigcap_{u\in U} [\![\phi]\!]_{\langle M[x:=u],f_{\mathcal{V}}\rangle}$ where $M[x := u]$ is a BDI structure obtained by replacing the interpretation of x in M with u
- $[\![\mathsf{pos}(e)]\!]_{\langle M,f_{\mathcal{V}}\rangle} = \{(w,t) \mid pos_{w,t} \ni e\}$
- $[\![\mathsf{X}^e({}_{r_1p_1}\,\phi_1 \mid \cdots \mid {}_{r_np_n}\,\phi_n)]\!]_{\langle M,f_{\mathcal{V}}\rangle} = \{(w,t) \mid$ for $i = 1,\cdots,n$, there are mutually disjoint sets $T_1,\cdots,T_n$ that satisfy the following$\}$
  - $\star$ $T_i \subseteqq \{t' \mid (w,t') \in [\![\phi_i]\!]_{\langle M,f_{\mathcal{V}}\rangle}\}$
  - $\star$ $(\sum_{t'\in T_i} R^e_w(t,t'))\ r_i\ p_i$ (note that each $r_1,\cdots,r_n$ is $\geqq$ or $>$)
- $[\![\mathsf{BEL}^a({}_{r_1p_1}\,\phi_1 \mid \cdots \mid {}_{r_np_n}\,\phi_n)]\!]_{\langle M,f_{\mathcal{V}}\rangle} = \{(w,t) \mid$ for $i = 1,\cdots,n$, there are mutually disjoint sets $W_1,\cdots,W_n$ that satisfy the following$\}$
  - $\star$ $W_i \subseteqq \{w' \mid (w',t) \in [\![\phi_i]\!]_{\langle M,f_{\mathcal{V}}\rangle}\}$
  - $\star$ $(\sum_{w'\in W_i} \mathcal{B}^t_a(w,w'))\ r_i\ p_i$
- The same goes for DESIRE^a and INTEND^a
- If $\mathfrak{X} \in \mathcal{V}$, $[\![\mathfrak{X}]\!]_{\langle M,f_{\mathcal{V}}\rangle} = f_{\mathcal{V}}(\mathfrak{X})$

Accordingly, a formula ϕ, with (or without) free occurrences of a formula variable symbol $\mathfrak{X}$, can be regarded as a function $f_\phi : Swt \to Swt$, which receives an interpretation of $\mathfrak{X}$ as an argument and returns an interpretation of ϕ. Therefore, we define that

- $[\![\mu\mathfrak{X}.\phi]\!]_{\langle M,f_{\mathcal{V}}\rangle}$ is the least fixed-point of f_ϕ.

Here, the least fixed-point exists since f_ϕ in this case is monotonic by definition[6].

From the properties of $\mathcal{B}^t_a$, $\mathcal{D}^t_a$ and $\mathcal{I}^t_a$, the introspection axioms hold; i.e. if Op is one of BEL^a, DESIRE^a or INTEND^a, then $\mathsf{Op}({}_{r_1p_1}\,\phi_1 \mid \cdots \mid {}_{r_np_n}\,\phi_n) \Leftrightarrow \mathsf{BEL}^a\,\mathsf{Op}({}_{r_1p_1}\,\phi_1 \mid \cdots \mid {}_{r_np_n}\,\phi_n)$ is valid.

C. Deduction system

In this section we describe the deduction system of $\mathcal{TOMATOes}$-$\mathcal{P}$ using sequent calculus.

We identify α-equivalent formulas. We regard the left side of "$\to$" of a sequent as a (finite) multi-set of formulas, and likewise for the right side (thus we do not have the exchange rule). Hereafter, we sometimes enclose a whole sequent in [ ] to clarify the range of the sequent in the text.

We will use capital Greek letters (Σ, Δ etc.; including letters with a hash such as Σ', and Δ') to denote multi-sets of 0 or more formulas.

1) Inference rules: Now let us enumerate the inference rules. Note that there are two more rules described in Sec. II-C2.

In the $\forall$L rule, t is an arbitrary term. In the $\forall$R rule, y is a variable symbol that does not occur freely in the conclusion of the rule. In the evAll rule, $\{e_1,\cdots,e_n\}$ is equivalent to $\mathcal{E}$.

In the $\mathsf{Op}_{\text{excl}}$, $\mathsf{Op}_{\geqq}$R and $\mathsf{Op}_{>}$R rules, Op may be one of X^e, BEL^a, DESIRE^a or INTEND^a (and must be same for one application of this rule). $\mathsf{Op}_{\text{excl}}$ rule means that any subformula of the form shown in the assumption anywhere in the sequent can be replaced by the formula shown in the conclusion. In this rule, $n \geqq 2$, and for $i = 1,\cdots,n$, ψ_i is $\mathfrak{X}_i \wedge \bigwedge_{1\leqq j\leqq n, i\neq j} \neg\mathfrak{X}_j$ where $\mathfrak{X}_1,\cdots,\mathfrak{X}_n$ are formula variable symbols that do not freely occur in the conclusion of the rule. This rule is provided so that we can decompose formulas in the form of $\mathsf{Op}(\cdots|\cdots)$ into those in the form of $\mathsf{Op}\,{}_{r_1p_1}\,\phi$, by inversely applying it.

2) Additional inference rules: Here we set Op to be one of X^e, BEL^a, DESIRE^a or INTEND^a. Let $\Gamma = \{\mathsf{Op}\,{}_{r_1p_1}\,\psi_1,\cdots,\mathsf{Op}\,{}_{r_np_n}\,\psi_n\}$, where each $r_1,\cdots,r_n$ is $\geqq$ or $>$, and $\mathfrak{Q} = \{\psi_1,\cdots,\psi_n\}$.

We say that a set $Z = \{Q_1,\cdots,Q_m\}$ (where each Q_i is a subset of $\mathfrak{Q}$) is a satisfaction request set (SRS) of Γ iff the following m-variable linear simultaneous inequation has any solution.

$$\begin{cases} \sum_{1\leqq j\leqq m} x_j = 1 & \\ x_j > 0 & (\text{for } 1 \leqq j \leqq m) \\ (\sum_{1\leqq j\leqq m, \psi_i\in Q_j} x_j)\ r_i\ p_i & (\text{for } 1 \leqq i \leqq n) \end{cases}$$

Then, Γ is satisfiable iff there is an SRS Z of Γ where each element of Z are satisfiable. For example, if $\Gamma = \{\mathsf{X}^{e_1}_{\geqq 0.3}\,\psi_1, \mathsf{X}^{e_1}_{\geqq 0.4}\,\psi_2, \mathsf{X}^{e_1}_{\geqq 0.6}\,\psi_3\}$ and $Z = \{\{\psi_1,\psi_2\},\{\psi_3\}\}$, Z is an SRS of Γ. If $\{\psi_1,\psi_2\}$ and ψ_3 are both satisfiable, so is Γ.

If, for $Z, Z' \subset 2^{\mathfrak{Q}}$, some $Q \in Z$ and $Z'' \subset 2^Q$ exist and $Z' = (Z \cup Z'') \setminus \{Q\}$ holds, we write $Z \succ Z'$. We call Z an essential SRS (eSRS) of Γ if Z is an SRS of Γ and there is no SRS Z' of Γ which satisfies $Z \succ Z'$. It is easy to show that Γ is satisfiable iff there is an eSRS Z of Γ and each element of Z is satisfiable. In other words, Γ is unsatisfiable iff for any eSRS Z of Γ, there is an unsatisfiable element of Z.

Let $Z_1 = \{Q_{1,1},\cdots,Q_{1,m_1}\},\cdots,Z_k = \{Q_{k,1},\cdots,Q_{k,m_k}\}$ be the enumeration of all eSRSs of Γ. Then for any sequence of positive integers $j_1,\cdots,j_k$, where $1 \leqq j_1 \leqq m_1$, $\cdots, 1 \leqq j_k \leqq m_k$, the following is an inference rule of $\mathcal{TOMATOes}$-$\mathcal{P}$, provided that Op is one of X^e, DESIRE^a or INTEND^a.

$$\frac{Q_{1,j_1} \to \quad \cdots \quad Q_{k,j_k} \to}{\Gamma \to}\ \mathsf{XDI\text{-}KD}$$

If Op is BEL^a, the following is an inference rule. Here Σ is a multi-set of formulas whose top-level operator is DESIRE^a or INTEND^a (with the same agent a). Γ and Σ remain in the assumption so as to make the introspection axioms provable.

$$\frac{\Gamma, Q_{1,j_1}, \Sigma \to \quad \cdots \quad \Gamma, Q_{k,j_k}, \Sigma \to}{\Gamma, \Sigma \to}\ \mathsf{BEL\text{-}KD45\text{-}DI}$$

If we want to consider the realism property, we have to modify these rules. It will be our future work.

3) Provability: A sequent S is said to be derivable from a set L of sequents if $S \in L$, or there is an inference rule $\frac{S_1,\cdots,S_n}{S}$ ($n \geqq 0$) and all of $S_1,\cdots,S_n$ are derivable from L.

We say that a sequent S is provable if one of the following conditions is satisfied. Here $\phi^n(\mathfrak{X})$ is defined as $\phi^0(\mathfrak{X}) = \mathfrak{X}$ and $\phi^n(\mathfrak{X}) = \phi[\mathfrak{X} := \phi^{n-1}(\mathfrak{X})]$.

1) S is derivable from $\emptyset$.

$$\frac{}{\phi \to \phi}\ \text{Initial} \quad \frac{\Sigma \to \Delta}{\Sigma, \Sigma' \to \Delta, \Delta'}\ \text{Weak} \quad \frac{\Sigma, \phi, \phi \to \Delta}{\Sigma, \phi \to \Delta}\ \text{CL} \quad \frac{\Sigma \to \Delta, \phi, \phi}{\Sigma \to \Delta, \phi}\ \text{CR} \quad \frac{\Sigma \to \Delta, \phi}{\Sigma, \neg\phi \to \Delta}\ \neg\text{L} \quad \frac{\Sigma, \phi \to \Delta}{\Sigma \to \Delta, \neg\phi}\ \neg\text{R} \quad \frac{\Sigma, \phi \to \Delta \quad \Sigma, \psi \to \Delta}{\Sigma, \phi \vee \psi \to \Delta}\ \vee\text{L}$$

$$\frac{\Sigma \to \Delta, \phi, \psi}{\Sigma \to \Delta, \phi \vee \psi}\ \vee\text{R} \quad \frac{\Gamma, \phi[\mathfrak{X} := \mu\mathfrak{X}.\phi] \to \Delta}{\Gamma, \mu\mathfrak{X}.\phi \to \Delta}\ \mu\text{L} \quad \frac{\Gamma \to \Delta, \phi[\mathfrak{X} := \mu\mathfrak{X}.\phi]}{\Gamma \to \Delta, \mu\mathfrak{X}.\phi}\ \mu\text{R} \quad \frac{\Gamma, \mathsf{Op}_{\geqq 1-p}\, \neg\phi \to \Delta}{\Gamma \to \Delta, \mathsf{Op}_{\geqq p}\, \phi}\ \mathsf{Op}_{\geqq}\text{R} \quad \frac{\Gamma, \mathsf{Op}_{>1-p}\, \neg\phi \to \Delta}{\Gamma \to \Delta, \mathsf{Op}_{>p}\, \phi}\ \mathsf{Op}_{>}\text{R}$$

$$\frac{\Sigma, \phi[x := t] \to \Delta}{\Sigma, \forall x\phi \to \Delta}\ \forall\text{L} \quad \frac{\Sigma \to \Delta, \phi[x := y]}{\Sigma \to \Delta, \forall x\phi}\ \forall\text{R} \quad \frac{}{\to \mathsf{pos}(e_1), \cdots, \mathsf{pos}(e_n)}\ \mathsf{evAll} \quad \frac{\cdots\ \mathsf{Op}_{r_1 p_1}(\phi_1 \wedge \psi_1) \wedge \cdots \wedge \mathsf{Op}_{r_n p_n}(\phi_n \wedge \psi_n)\ \cdots}{\cdots\ \mathsf{Op}(_{r_1 p_1}\, \phi_1 \mid \cdots \mid {}_{r_n p_n}\, \phi_n)\ \cdots}\ \mathsf{Op}_{\text{excl}}$$

Fig. 1. Inference rules of $\mathcal{TOMASTOES}$-$\mathcal{P}$

2) $S = [\Sigma, \mu\mathfrak{X}.\phi \to \Delta]$ where $\mathfrak{X}$ does not occur freely in Σ, Δ, and there is a positive integer n s.t. $[\Sigma, \phi^n(\mathfrak{X}) \to \Delta]$ is derivable from $\{[\Sigma, \mathfrak{X} \to \Delta]\}$.

The soundness of $\mathcal{TOMASTOES}$-$\mathcal{P}$ can be proved just like in [3] and is relatively easy. We also expect that we can show the completeness of $\mathcal{TOMASTOES}$-$\mathcal{P}$. We guess that the proof is similar to the sketch we gave in our previous study[5].

III. Modeling Werewolf

In this section, as an example of actual Werewolf play, we select a story written in [7] and show some descriptions and proofs, though rather simple, about the game process near the end of the story. The story is a description of a Werewolf match between AIs, and the AI Werewolf Project submitted this story to the 3rd Hoshi-Shin'ichi award of Nikkei Inc. as "a novel that an AI was involved in writing". However, to make the story in the style of SF, the terminology of the Werewolf game was replaced with other words; e.g. "werewolf" was changed into "AI", "seer" into "analyst", etc. In the following, we turn those words back into the original ones. Note that the spoiler of the story is included hereafter.

A. Representing some situation

Near the final stage of the game, five players (out of ten) are still alive; player1 (medium), player2 (confessed[2] to being a seer, but is actually a lunatic[3]), player4 (villager), player6 (werewolf) and player9 (villager). As the game progresses, the players come to believe that there is only one werewolf alive. Though this belief is actually uncertain, for simplicity, we will assume that the players are certain in this belief.

One day, player2 says that (s)he has divined player4 and that player4 is a werewolf. However, in the past, player5 (who has confessed to being a seer, and was already executed) said that player4 is a human.

So the village has to make a decision whether to believe player2 as a seer and execute player4 or to doubt player2 as a werewolf and execute him/her. We can describe this situation as follows. AX operator denotes the next day, and $executed$ denotes "executed yesterday".

$$\mathsf{BEL}^a(seer(player2) \supset \mathsf{AX}(executed(player4) \supset village\_win)) \quad (1)$$

$$\mathsf{BEL}^a(wolf(player2) \supset \mathsf{AX}(executed(player2) \supset village\_win)) \quad (2)$$

$$\mathsf{BEL}^a_{\geqq 0.8}(lunatic(player2) \supset \mathsf{AX}(\neg executed(player2) \supset village\_lose)) \quad (3)$$

[2]This action is also called "coming out".

[3]Also called "possessed".

$$\dfrac{\dfrac{\dfrac{\dfrac{\phi \to \phi \quad \psi \to \psi}{\phi \supset \psi, \neg\psi, \neg\neg\phi \to}\ \text{classic}}{\mathsf{BEL}^a_{\geqq .9}(\phi \supset \psi), \mathsf{BEL}^a\, \neg\psi, \mathsf{BEL}^a_{>.1}\, \neg\neg\phi \to}\ \begin{matrix}\mathsf{BEL\text{-}KD45\text{-}DI} \\ +\ \text{Weak}\end{matrix}}{\mathsf{BEL}^a_{\geqq .9}(\phi \supset \psi), \mathsf{BEL}^a\, \neg\psi \to \mathsf{BEL}^a_{\geqq .9}\, \neg\phi}\ \mathsf{Op}_{\geqq}\text{R}}{\to \mathsf{BEL}^a_{\geqq .9}(\phi \supset \psi) \wedge \mathsf{BEL}^a\, \neg\psi \supset \mathsf{BEL}^a_{\geqq .9}\, \neg\phi}\ \text{classic}$$

Fig. 2. Proof of formula (4)

Formula (3) represents that if player2 is a lunatic, then (since the werewolf is very likely to know it) unless (s)he is executed, tomorrow the werewolf will collude with him/her and the villagers' side will likely lose. 0.8 in that formula may not be accurately 0.8, but will be some relatively large value.

B. Inferences using mental state operators with probability

At this time, player6 says that "if player2 were a werewolf, a lunatic would very likely support a werewolf in some way, but actually no such action has been observed". Viewing this speech as a decisive factor, many players vote for player4, and as a result, player4 is executed. Here we write "player x has acted by supporting a werewolf in some way" as $support(x)$. Accordingly, "since no support action by a lunatic has been observed, maybe player2 is not a werewolf" is represented as follows.

$$\mathsf{BEL}^a_{\geqq .9}\,(wolf(player2) \supset \exists x\, support(x)) \wedge \mathsf{BEL}^{\bar{a}}\, \neg\exists x\, support(x) \supset \mathsf{BEL}^a_{\geqq .9}\, \neg wolf(player2) \quad (4)$$

This formula can be proved as in Fig. 2. Here, for sake of readability, we have replaced some subformulas with ϕ, ψ etc. and displayed some applications of rules in a bundled manner (the same goes for the proofs hereafter). The application of BEL-KD45-DI depends on the fact that the only eSRS of $\{\mathsf{BEL}^a_{\geqq .9}\,(\phi \supset \psi), \mathsf{BEL}^a_{\geqq 1}\, \neg\psi, \mathsf{BEL}^a_{>.1}\, \neg\neg\phi\}$ is $\{\{\phi \supset \psi, \neg\psi, \neg\neg\phi\}\}$.

In this game, most players commonly have the belief that either player2 or player5, each of which has confessed to being a seer, is highly likely to be a werewolf. As a result, at this time most players apparently think that player2 is either a seer or a werewolf. Given this belief, together with the result from formula (4), one can conclude that player2 is likely to be a seer. This process can be represented as follows and is proved in Fig. 3.

$$\mathsf{BEL}^a_{\geqq .8}\,(wolf(player2) \vee seer(player2)) \wedge \mathsf{BEL}^{\bar{a}}_{\geqq .9}\, \neg wolf(player2) \supset \mathsf{BEL}^a_{\geqq .7}\, seer(player2) \quad (5)$$

$$\dfrac{\dfrac{\dfrac{\dfrac{\phi \to \phi \quad \psi \to \psi}{\phi \vee \psi, \neg\phi, \neg\psi \to}\ \text{classic}}{\mathsf{BEL}^a_{\geqq .8}(\phi \vee \psi), \mathsf{BEL}^a_{\geqq .9}\,\neg\phi, \mathsf{BEL}^a_{>.3}\,\neg\psi \to}\ \begin{array}{c}\text{BEL-KD45-DI}\\ +\,\text{Weak}\end{array}}{\mathsf{BEL}^a_{\geqq .8}(\phi \vee \psi), \mathsf{BEL}^a_{\geqq .9}\,\neg\phi \to \mathsf{BEL}^a_{\geqq .7}\,\psi}\ \mathsf{Op}_{\geqq}\mathrm{R}}{\to \mathsf{BEL}^a_{\geqq .8}(\phi \vee \psi) \wedge \mathsf{BEL}^a_{\geqq .9}\,\neg\phi \supset \mathsf{BEL}^a_{\geqq .7}\,\psi}\ \text{classic}$$

Fig. 3. Proof of formula (5)

In addition, we can write some of the strategies for vote of the players on the villagers' side (if determining which player to execute is directly linked to winning or losing the next day, vote for that player) as follows. Here $\mathsf{pos\_only}(e)$ is the abbreviation for $\mathsf{pos}(e) \wedge \bigwedge_{e' \neq e} \neg\mathsf{pos}(e')$ and denotes "choose e as the action". In this case, we can prove (in a similar way to the proofs described above, though the proof is somewhat long) that if $\mathsf{BEL}^a_{\geqq .7}\, seer(player2)$ and formulas (1), (6) hold, then $\mathsf{INTEND}^a_{\geqq .7}\,\mathsf{pos\_only}(vote\_for\_player4)$ holds. In addition, if $\mathsf{BEL}^a_{\geqq .05}\, wolf(player2)$ also holds, then $\mathsf{INTEND}^a_{\geqq .05}\,\mathsf{pos\_only}(vote\_for\_player2)$ is inferred. In this way, we can express the situation that if someone cannot get a certain belief then (s)he can hardly decide on a single intention. In the original BDI model, such situation cannot be expressed (an agent always chooses a single intention for one goal at a time).

$$\mathsf{BEL}^a_{\geqq p}\ \mathsf{AX}(executed(playerN) \supset village\_win) \supset \mathsf{INTEND}^a_{\geqq p}\,\mathsf{pos\_only}(vote\_for\_playerN) \quad (6)$$

$$\mathsf{BEL}^a_{\geqq p}\ \mathsf{AX}(\neg executed(playerN) \supset village\_lose) \supset \mathsf{INTEND}^a_{\geqq p}\,\mathsf{pos\_only}(vote\_for\_playerN) \quad (7)$$

In this game, as player2 is actually a lunatic, the villagers' side loses in the course of formula (3). If there was a player who emphasized the possibility that player2 was a lunatic and had a belief like $\mathsf{BEL}^a_{<.5}(wolf(player2) \vee seer(player2))$, then through a process like formula (3), the chances that player2 is a seer would be low for him/her, and (s)he would have chosen a different way of voting.

C. Nested mental state operators

A belief of an agent that "if player2 is a lunatic, then the werewolf is very likely to know it", which was previously referred to, can be represented as formula (8) by using nested mental state operators. Here σ is $\forall x(wolf(x) \supset \mathsf{BEL}^x\, lunatic(player2))$. In addition, a belief that "the next day, if σ holds, then villagers' side will lose unless player2 is executed" can be represented as formula (9). We can prove that (3) can be inferred from formula (8) and (9). We will omit the details of the proof, but some small fragments of the proof is shown in Fig. 4 (where $\mathcal{E} = \{e_1, \cdots, e_n\}$).

$$\mathsf{BEL}^a_{\geqq .8}\,(lunatic(player2) \supset \mathsf{AG}\,\sigma) \quad (8)$$

$$\mathsf{BEL}^a\,\mathsf{AX}(\sigma \wedge \neg executed(player2) \supset village\_lose) \quad (9)$$

It is also possible to handle an agent's desires and intentions about other agents' mental states. Here, if the werewolf actually knows that player2 is a lunatic, (s)he can lead the game into a win for the werewolves by making other players not vote for player2. In the situation described above, to do so, one can state that player2 is not a werewolf. This can

$$\dfrac{\dfrac{\dfrac{\Sigma, \xi, \mathsf{AX}\,\mathsf{AG}\,\xi \to \Delta}{\Sigma \to \Delta, \neg\xi \vee \neg\,\mathsf{AX}\,\mathsf{AG}\,\xi}\ \text{classic}}{\Sigma \to \Delta, \mu\mathfrak{X}.(\neg\xi \vee \neg\,\mathsf{AX}\,\neg\mathfrak{X})}\ \mu\mathrm{R}}{\Sigma, \mathsf{AG}\,\xi \to \Delta}\ \text{classic}$$

$$\dfrac{\dfrac{\dfrac{\xi_1, \xi_2 \to \zeta}{\xi_1, \xi_2, \neg\zeta \to}\ \neg\mathrm{L}}{\mathsf{AX}^{e_1}\xi_1, \mathsf{AX}^{e_1}\xi_2, \mathsf{X}^{e_1}_{>0}\neg\zeta \to}\ \mathsf{XDI\text{-}KD} \quad \cdots \quad \dfrac{\dfrac{\xi_1, \xi_2 \to \zeta}{\xi_1, \xi_2, \neg\zeta \to}\ \neg\mathrm{L}}{\mathsf{AX}^{e_n}\xi_1, \mathsf{AX}^{e_n}\xi_2, \mathsf{X}^{e_n}_{>0}\neg\zeta \to}\ \mathsf{XDI\text{-}KD}}{\mathsf{AX}\,\xi_1, \mathsf{AX}\,\xi_2 \to \mathsf{AX}\,\zeta}\ \mathsf{Op}_{\geqq}\mathrm{R}\ \text{etc.}$$

Fig. 4. Parts of proof of formula (8) $\wedge$ (9) $\supset$ (3)

explain the behavior of player6 in the actual play. Formula (10) represents this behavior.

$$\mathsf{DESIRE}^{player6}\ \forall x(human(x) \supset \mathsf{BEL}^x\,\neg wolf(player2)) \quad (10)$$

D. State transition

Here we abbreviate $\mu\mathfrak{X}.(\bigwedge_{a\in\mathcal{A}}\mathsf{BEL}^a(\phi \wedge \mathfrak{X}))$ as $\mathsf{M\text{-}BEL}\phi$ (ϕ is a mutual belief among all agents) and $\mu\mathfrak{X}.(\phi \wedge \psi \vee \neg\psi \wedge \mathsf{AX}\,\mathfrak{X})$ as $\mathsf{A}(\phi\,\mathsf{N}\,\psi)$ (ϕ holds the next time ψ holds; N is regarded as the "atnext" operator).

Thus far, we have used the AX operator to denote "the next day". However, to capture the process of modifying beliefs by agents' speech acts, a state transition caused by each speech act has to be described. So, from here on, we will use the AX operator to denote the progress of one step in the game, e.g. one speech act of an agent, vote for execution, start of a new day, etc. In this case, "ϕ will hold the next day" can be written as $\mathsf{A}(\phi\,\mathsf{N}\,daystart)$ where $daystart$ is "a new day has just started".

We treat speech acts as events. However, syntactically, the set of events has to be finite (Sec. II-A). For this reason (and for simplicity), we choose some sufficiently large finite set SA of formulas, and limit the set of the contents of the speech acts to SA.

Again for simplicity, we will consider only the "inform" type of speech acts defined in FIPA[8]. We write "agent a says ϕ ($\in SA$) to the set of agents A" as $inform(a, A, \phi)$ and treat this as an event. (From the aspect of the implementation, a logic programming language such as Prolog often does such a thing, i.e. treats a formula as data.) If $A = \mathcal{A}$ (the set of all agents), we simply write $inform(a, \phi)$ (this is a common case in the Werewolf game). We assume that each agent works so as to satisfy the following formula (where $done(e)$ denotes "event e has just occurred"):

$$\bigwedge_{inform(a,\phi)\in\mathcal{E}}(\mathsf{pos}(inform(a, \phi)) \supset \mathsf{AX}^{inform(a,\phi)}\,\mathsf{M\text{-}BEL}\,done(inform(a, \phi))) \quad (11)$$

i.e. when someone says something, the fact that (s)he says so becomes a common belief. We also assume that the following formula holds; when an agent a intends to do the action e, then e is actually executed. This is regarded as a basic property of BDI agents.

$$\bigwedge_{e\in\mathcal{E}}(\mathsf{INTEND}^a\,\mathsf{pos\_only}(e) \supset \mathsf{pos\_only}(e)) \quad (12)$$

When an agent hears someone say something, (s)he decides whether to modify his/her belief in his/her own way. Here

$$\dfrac{\dfrac{\dfrac{\bigwedge_{a\in\mathcal{A}} \mathsf{BEL}^a(done(\epsilon) \wedge \mathsf{M\text{-}BEL}done(\epsilon)) \rightarrow \mathsf{BEL}^a\, done(\epsilon)}{\mathsf{M\text{-}BEL}done(\epsilon) \rightarrow \mathsf{BEL}^a\, done(\epsilon)}\ \mu\mathrm{L}}{\mathsf{AX}^\epsilon\, \mathsf{M\text{-}BEL}done(\epsilon) \rightarrow \mathsf{AX}^\epsilon\, \mathsf{BEL}^a\, done(\epsilon)}\ \text{XDI-KD etc.}}{\mathsf{pos\_only}(\epsilon), \mathsf{pos}(\epsilon) \supset \mathsf{AX}^\epsilon\, \mathsf{M\text{-}BEL}done(\epsilon) \rightarrow \mathsf{AX}\, \mathsf{BEL}^a\, done(\epsilon)}\ \text{classic etc.}$$

Fig. 5. Part of proof of formula appeared in Sec. III-D

we assume that when player6 says $(\phi \supset \psi) \wedge \neg\psi$ (where $\phi \equiv wolf(player2)$ and $\psi \equiv \exists x\, support(x)$), some player a accepts $\phi \supset \psi$ with certainty level $\geqq 0.9$ and $\neg\psi$ with certainty 1. That is, a runs so as to satisfy formula (13) described below, where ϵ is an event $inform(player6, (\phi \supset \psi) \wedge \neg\psi)$.

$$\mathsf{BEL}^a\, done(\epsilon) \supset \mathsf{BEL}^a_{\geqq .9}(\phi \supset \psi) \wedge \mathsf{BEL}^a\, \neg\psi \tag{13}$$

$$\forall x \forall t(\mathsf{BEL}^a_{rp}\, wolf(x) \wedge mod\_bel \wedge time(t) \supset \mathsf{A}(\mathsf{BEL}^a_{rp}\, wolf(x)\, \mathsf{U}\, \rho)) \tag{14}$$

where $\rho \equiv (mod\_bel \wedge \exists t'(time(t') \wedge t < t'))$

Additionally, we assume that a runs so as to satisfy also formula (14) described above, where $time$ is a system predicate to get the current time, and that when a wants to modify his/her belief about the werewolf because (s)he has gotten a new information, (s)he makes mod_bel hold. Here r is an arbitrary comparison operator and p is an arbitrary element of $[0, 1]$. In other words, an estimate about the werewolf is retained over time until a attempts to modify it by making mod_bel hold.

Suppose that a is neutral about the success or failure of ϕ, i.e. $\mathsf{BEL}^a_{=.5}\, \phi$. If player6 satisfies θ and $\theta \supset \mathsf{INTEND}^{player6}\, \mathsf{pos\_only}(\epsilon)$ where θ is formula (10) and ϵ is an event described above, then the event ϵ occurs. Accordingly, $\mathsf{BEL}^a_{\leqq .1}\, \phi$ holds the next time (by formula (4)), and if a makes mod_bel hold at the same time, this belief is retained until another estimation is obtained. Regarding these beliefs as states, we can express the process of state transitions caused by another agent's speech act in this way.

We can prove $\theta \wedge (\theta \supset \mathsf{INTEND}^{player6}\, \mathsf{pos\_only}(\epsilon)) \wedge$ formula (12) $\wedge$ (11) $\wedge\, \mathsf{AG}$ (13) $\supset \mathsf{AX}\, \mathsf{BEL}^a_{\leqq .1}\, \phi$. The proof of this formula is again omitted, but we show a small fragment of the proof in Fig. 5. We can also prove $\mathsf{AX}\, \mathsf{BEL}^a_{\leqq .1}\, \phi \wedge \mathsf{AG}$ (formula (14)) $\supset \mathsf{AX} \forall t(mod\_bel \wedge time(t) \supset \mathsf{A}(\mathsf{BEL}^a_{\leqq .1}\, \phi\, \mathsf{U}\, \rho))$.

IV. Discussion

In this paper, we only showed that an inference rule can be used to prove some properties. In general, the existence of a deduction system does not mean the existence of an algorithm which infers goals efficiently. To create a logic-based practical AI Werewolf engine, we need to devise an algorithm to generate goal candidates to be inferred and infer those goals in a realistic amount of time. It might be desirable to have an ability to exclude goals which are not likely to hold from the candidates, instead of ensuring completeness.

There have already been a number of attempts to realize an AI Werewolf; some of them use machine learning. For example, [9] uses SVM to estimate werewolves. However, one can imagine that the best part of the Werewolf game is enjoying the process of logical thinking rather than being concerned about winning or losing. Given such a stance, one can expect that if a logic-based AI Werewolf engine with enough skill can be realized, by outputting its thought processes and decision making during the game, we can enjoy looking back on episodes in the game after it has finished.

Apart from the application to the Werewolf game, there have been studies on extended BDI model with probabilistic mental states. For instance, Ma et al.[10] propose a framework for probabilistic plan selection under beliefs with uncertainty. Coelho et al.[11] proposes an integration of symbolic and probabilistic approaches for agent programming on Jason[12], an interpreter and development environment for AgentSpeak(L). However, unlike these studies, our approach is logic-based and thinking processes including probabilities are directly expressed as deductions. This feature is considered to be suitable for applications that require logical thought.

V. Conclusion

Using an extended BDI logic with probabilistic mental states, we attempted to describe the processes of decision making with probabilistic intentions, and prove certain properties of players in the Werewolf game. Our future work will include developing an algorithm to generate goal candidates to be inferred and to infer those goals in a realistic amount of time, as described in Sec. IV, with the ultimate goal being the realization of a logic-based Werewolf agent. From the aspect of logic systems, we should also take into account the "realism" property ($\rightarrow$ Sec. II) and prove the completeness of $\mathcal{TOMATOes}$-$\mathcal{P}$.

References

[1] H. Osawa, F. Toriumi, M. Inaba, D. Katagami, K. Kajiwara, and K. Shinoda, "Agent's reasoning model for achieving aiwolf," in *Proc. of 19th Game Programming Workshop*, 2014, pp. 157–161, (In Japanese).

[2] N. Nide and S. Takata, "Stochastic strategy of agents with uncertain beliefs and BDI model," in *Proc. of 30th Annual Conference of JSAI*, 2016, (In Japanese).

[3] N. NIDE, S. Takata, and M. Fujita, "BDI logic with probabilistic transition and fixed-point operator," in *Proc. of CLIMA '09*, 2009, pp. 71–86.

[4] A. S. Rao and M. P. Georgeff, "Modeling rational agents within a BDI-architecture," in *Readings in Agents*, M. N. Huhns and M. P. Singh, Eds. Morgan Kaufmann, San Francisco, 1997, pp. 317–328.

[5] N. Nide, S. Takata, and M. Fujita, "Modeling cooperative actions using an extended BDI logic $\mathcal{TOMATOes}$," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 26, no. 1, pp. 13–24, 2011, (In Japanese).

[6] A. Tarski, "A lattice-theoretical fixpoint theorem and its application," *Pacific Journal of Mathematics*, vol. 5, pp. 285–309, 1955.

[7] AI Werewolf Project, "Are you an AI? TYPE-S," `http://aiwolf.org/control-panel/wp-content/uploads/2016/03/YASAI20150930_Short.pdf`, 2016, (In Japanese).

[8] IEEE Foundation for Intelligent Physical Agents, "FIPA communicative act library specification," `http://www.fipa.org/specs/fipa00037/SC00037J.html`, 2002.

[9] K. Kajiwara, F. Toriumi, M. Inaba, H. Osawa, D. Katagami, K. Shinoda, H. Matsubara, and Y. Kano, "Development of AI wolf agent using SVM to detect werewolves," in *Proc. of 30th Annual Conference of JSAI*, 2016, (In Japanese).

[10] J. Ma, W. Liu, J. Hong, L. Godo, and C. Sierra, "Plan selection for probabilistic BDI agents," in *Proc. of ICTAI 2014*, 2014, pp. 83–90.

[11] F. Coelho and V. Nogueira, "Probabilistic selection in AgentSpeak(L)," *Computer Research Repository*, vol. abs/1409.3717, 2014.

[12] R. H. Bordini, J. F. Hübner, and M. Wooldridge, *Programming Multi-Agent Systems in AgentSpeak using Jason*. John Wiley & Sons, 2007.