The Semantics of Potential Intentions

Xiaocong Fan and **John Yen**
School of Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802
{zfan, jyen}@ist.psu.edu

Abstract

The SharedPlans theory provides an axiomatic framework of collaborative plans based on four types of intentional attitudes. However, there still lacks an adequate semantics for the 'potential intention' operators. In this paper, we give a formal semantics to potential intentions, and examine models that can validate various relations between beliefs, intentions, and potential intentions.

Introduction

Philosophers have long struggled over how best to characterize the concept of intention (Searle 1983; Bratman 1987), which is intimately connected with means-ends reasoning. Normal modal logics have been employed to define possible-worlds semantics for intentions (Cohen & Levesque 1990; Rao & Georgeff 1995). Konolige & Pollack (1993) provide a representationalist theory of intention, using cognitive structures to directly represent intentions and the means-ends relationship among intentions. Singh & Asher (1993) give a theory of intentions based on Discourse Representation Theory. Existing solutions have been extended to investigate intentions involving groups and cooperations (Grosz & Sidner 1990; Singh 1993; Herzig & Longin 2002).

The SharedPlans theory (Grosz & Kraus 1996) provides an axiomatic framework of collaborative plans based on four types of intentional attitudes. Operators Int.To and Int.Th represent intentions that have been adopted by an agent, while Pot.Int.To and Pot.Int.Th represent potential intentions—intentions that an agent would like to adopt, but to which it is not yet committed. Int.To and Pot.Int.To apply to actions while Int.Th and Pot.Int.Th apply to propositions.

However, Grosz & Kraus only informally characterized what it means for an agent to have a potential intention. There still lacks an adequate semantics for the 'potential intention' operators. Our aim in this paper is to present a formal semantics of potential intentions and investigate the relationships of potential intentions with agent beliefs and normal intentions.

Desire is a potential influencer of conduct while intention is a conduct-controlling pro-attitude (Bratman 1990). From such a sense, potential intentions are agent desires. Desires play an important role in determining goals and intentions that stand for the consistent worlds an agent is striving for (Cohen & Levesque 1990). The notion of 'potential intentions' differs from the concepts of 'goal' and 'want' as appeared in the literature. Goals are *chosen* desires (Cohen & Levesque 1990). Thus, one major difference between potential intentions and goals is that an agent cannot hold incompatible goals but can hold incompatible potential intentions. Potential intentions as used in the SharedPlans theory are also different from the 'want' attitude proposed by Sadek (1992). $want(A, p)$ abbreviates $Bel(A, p) \vee Int(A, Bel(A, p))$. Thus, an agent wants what it believes. This is quite different from potential intentions.

The paper is organized as follows. We introduce the motivations in Sec. 2. We then give the formal language (Sec. 3), model (Sec. 4), and semantics (Sec. 5), and examine models (Sec. 6) that can validate various relations between beliefs, intentions, and potential intentions.

Motivations

Possible worlds have been treated as collections of propositions (Halpern & Moses 1992), time lines representing a sequence of events (Cohen & Levesque 1990), and time trees with branching futures (Singh & Asher 1993; Rao & Georgeff 1995). A common ground of these approaches is that the accessibility relations are maps from a single world to a set of possible worlds. One exception in Rao & Georgeff's work (1995) is that they examined the relationship between the belief-, desire-, and intention-accessible worlds with respect to the *structure* of possible worlds, where a world is a *sub-world* of another if the former contains fewer paths but they are otherwise identical to each other.

From a representationalist's perspective, Konolige & Pollack (1993) used the notion of embedding graph to represent the structure of intentions by means of relations among relevant scenarios (i.e., a set of possible worlds) [1]. Singh & Asher (1993) employed an embedding function to assign strategies (i.e., the abstract specifications of the behavior of an agent. This is similar to recipes (Grosz & Sidner 1990)

[1] Rao & Georgeff (1995) examined the set relationship among the belief-, desire-, and intention-accessible worlds. This is different from the set relationship considered by Konolige & Pollack, who focused on the relationship between relevant intentions

and action expressions (Cohen & Levesque 1990)) to agents at different worlds and times.

Due to the unique nature of potential intentions and their relations with normal intentions, we choose to give a non-normal possible-worlds semantics to intentions and potential intentions, drawing upon techniques used by representationalist. Particularly, our model will involve three kinds of relationship among possible worlds: (a) pointwise relationship: from a world to a sub-world; (b) accessibility relationship: from a world to a set of worlds; and (c) set relationship: from a set of worlds to another set. Our model will also include an embedding function for assigning recipes to agents.

The dynamics of intentions considered here is based on the following observations on the SharedPlans theory:

(a) An agent may hold conflicting potential intentions but cannot hold conflicting intentions. Conflicts can arise because an agent may only have partial plans, or an agent may only have incomplete picture of the plans being used by the other group members.

(b) Potential intentions can be upgraded to intentions, as long as it would not cause intention conflicts. But not all intentions are originated from potential intentions.

(c) Potential intentions typically arise from means-ends reasoning (cf. Axioms A5 and A6 (Grosz & Kraus 1996)). Potential intentions are thus connected with recipes. We consider three kinds of *originated-from* relations between potential intentions and intentions:
• An agent with an intention that p hold will adopt a potential intention to do action α if the agent has a recipe for α that can *lead to* p;
• An agent has to follow some recipe to pursue an intention. A recipe may contain choice points, each is associated with several alternative courses of action. These alternatives typically serve the same ends (e.g., to make p hold) but may have different requirements (both physically and epistemically) on the agent. However, due to uncertainty of the situation, the agent may not know which one works better when reaching a choice point. In such a case, the agent can adopt potential intentions rather than full-fledged intentions so that it can make a better choice in a later deliberation process;
• A group of agents with an intention to do a course of action may reach an action whose doer has not been resolved yet. To be helpful, each agent in the group who has the capability and capacity can adopt a potential intention to do the action.

Formal language

The formal language L is a multi-modal predicate calculus, augmented with temporal operators from the branching-time logic CTL* (Emerson 1990).

Definition 1 *The alphabet of* L *has the following symbols: a denumerable set $Pred$ of predicate symbols; a denumerable set of constant symbols $Const \supseteq Const_{Ag} \cup Const_{Ac} \cup Const_{Gr} \cup Const_{Re} \cup \{\text{'true'}\}$ where the mutually disjoint sets $Const_{Ag}$, $Const_{Gr}$, $Const_{Ac}$ and $Const_{Re}$ are constants for agents, agent groups, primitive act-types (the empty act $nil \in Const_{Ac}$) and recipes, respectively; a denumerable set of variable symbols $Var \supseteq Var_{Ag} \cup Var_{Gr} \cup Var_{Ac} \cup Var_{Re}$ where the mutually disjoint sets $Var_{Ag} = \{x, y, x_1, \cdots\}$, $Var_{Gr} = \{G, G', \cdots\}$, $Var_{Ac} = \{a, b, a', \cdots\}$ and $Var_{Re} = \{\gamma, \gamma', \cdots\}$ are variables for agents, agent groups, primitive act-types and recipes, respectively; the recipe refinement operator $\sqsupseteq$; the membership relation operator $\in$; the classical connectives $\vee$ ('or'), $\neg$ ('not'), and the universal quantifier $\forall$; the modal operators* Bel, Pot.Int.To, Pot.Int.Th, Int.To, Int.Th*; the temporal operators* X *(next),* U *(until), and path quantifier* E *(some path in the future); the action constructor symbols ';' (sequential) , '|' (choice), '?' (test), '!' (achieve), and '*' (iterative); and the punctuation symbols ')' , '(', ',' and '.'.*

A term is either a constant or a variable. The syntax of well-formed state formulas and path formulas is defined in Figure 1. Let AE be the set of well-formed action expressions, F_{state} be the set of well-formed state formulas, α, β, e, e' range over AE, and ϕ, ψ, ϕ' range over F_{state}.

$\langle$pred$\rangle$ ::= any element of $Pred$
$\langle$ag-term$\rangle$ ::= any element of $Term_{Ag}$
$\langle$ac-term$\rangle$ ::= any element of $Term_{Ac}$
$\langle$gr-term$\rangle$ ::= any element of $Term_{Gr}$
$\langle$re-term$\rangle$ ::= any element of $Term_{Re}$
$\langle$var$\rangle$ ::= any element of Var
$\langle$term$\rangle$ ::= any element of $\cup_{s\in\{Ag,Ac,Gr,Re\}} Term_k$
$\langle$ac-exp$\rangle$ ::= $\langle$ac-term$\rangle$ | $\langle$re-term$\rangle$
| $\langle$ac-exp$\rangle$; $\langle$ac-exp$\rangle$ | $\langle$ac-exp$\rangle'|'\langle$ac-exp$\rangle$
| $\langle$s-fmla$\rangle$? | $\langle$s-fmla$\rangle$! | $\langle$ac-exp$\rangle*$
$\langle$s-fmla$\rangle$ ::= ($\langle$pred$\rangle$ $\langle$term$\rangle, \cdots, \langle$term$\rangle$)
| $\neg\langle$s-fmla$\rangle$ | $\langle$s-fmla$\rangle \vee \langle$s-fmla$\rangle$
| $\forall\langle$var$\rangle \cdot \langle$s-fmla$\rangle$ | E$\langle$p-fmla$\rangle$
| $\langle$ag-term$\rangle \in \langle$gr-term$\rangle$ | $\langle$ac-term$\rangle \in \langle$re-term$\rangle$
| $\langle$re-term$\rangle \sqsupseteq \langle$re-term$\rangle$ | Bel($\langle$ag-term$\rangle, \langle$s-fmla$\rangle$)
| Int.To($\langle$ag-term$\rangle$ | $\langle$gr-term$\rangle, \langle$ac-exp$\rangle, \langle$s-fmla$\rangle$)
| Pot.Int.To($\langle$ag-term$\rangle$ | $\langle$gr-term$\rangle, \langle$ac-exp$\rangle, \langle$s-fmla$\rangle$)
| Int.Th($\langle$ag-term$\rangle$ | $\langle$gr-term$\rangle, \langle$s-fmla$\rangle, \langle$s-fmla$\rangle$)
| Pot.Int.Th($\langle$ag-term$\rangle$ | $\langle$gr-term$\rangle, \langle$s-fmla$\rangle, \langle$s-fmla$\rangle$)
$\langle$p-fmla$\rangle$ ::= $\langle$s-fmla$\rangle$| $\neg\langle$p-fmla$\rangle$ | $\langle$p-fmla$\rangle \vee \langle$p-fmla$\rangle$
| $\forall\langle$var$\rangle \cdot \langle$p-fmla$\rangle$ | X$\langle$p-fmla$\rangle$ | $\langle$p-fmla$\rangle$U$\langle$p-fmla$\rangle$

Figure 1: The syntax

To simplify the presentation, we omit the time arguments of the modal operators used in the SharedPlans theory. Bel(A, ϕ) represents agent A believes ϕ; Int.To(A, α, C) and Int.Th(A, ϕ, C) represents that agent A under context C intends to do α, intends that ϕ hold, respectively; Pot.Int.To(A, α, C) and Pot.Int.Th(A, ϕ, C) represents that agent A under context C potentially-intends to do α, potentially-intends that ϕ hold, respectively. We use Pot.Int.Tx(A, α, C) to refer to either Pot.Int.To(A, α, C) or Pot.Int.Th(A, α, C). Such used, α refers to either an action (expression) or a proposition. Similar applies to Int.Tx (but keep the 'x' fixed in a specific definition).

Other connectives and operators such as $\wedge$, $\rightarrow$, F(sometime in the future), A (all paths in the future), and

G (all times in the future), can be defined as abbreviations.

A recipe, $\gamma = \langle e_\gamma, \rho_\gamma \rangle$, is composed of an action expression $e_\gamma \in AE$ and a constraint $\rho_\gamma \in F_{state}$. The doing of the course of action in e_γ under constraint ρ_γ constitutes the performance of γ.

Definition 2 *The roles contained in an action expression e, $Role(e)$, is defined recursively:*
0. $Role(nil) = \emptyset$; $Role(\phi?) = \emptyset$; $Role(\phi!) = \emptyset$;
1. $Role(a) = \{a\}$ if $a \in Term_{Ac}$;
2. $Role(\gamma) = Role(e_\gamma)$ if $\gamma \in Term_{Re}$;
3. $Role(e_1; e_2) = Role(e_1) \cup Role(e_2)$;
4. $Role(e_1|e_2) = Role(e_1) \cup Role(e_2)$;
5. $Role(e^*) = Role(e)$.

The Formal Model

The formal model treats each possible world as a time tree with a single past and a branching future (Singh & Asher 1993; Wooldridge & Jennings 1994; Rao & Georgeff 1995). Each world $w = \langle S_w, R_w \rangle$ is indexed with a set of states S_w partially ordered by temporal precedence R_w. In particular, we use w_0 to denote the starting state of world w (i.e., $\nexists s' \cdot s' R_w w_0$). A path in world w is any single branch of w starting from a given state and contains all the states in some linear sub-relation of R_w. A state transition is caused by the occurrence of a primitive action or an event.

Let the domain of quantification $D = D_{Ac} \cup D_{Re} \cup D_{Ag} \cup D_{Gr} \cup D_U$, where D_{Ac} is the set of all primitive act-types, D_{Re} is the set of recipes defined over D_{Ac}, D_{Ag} is the set of agents populating each world, $D_{Gr} = 2^{D_{Ag}} \setminus \emptyset$ is the set of agent groups, and D_U is the set of other individuals. Let $D_A = D_{Ag} \cup D_{Gr}$, and $D_T = D_{Ac} \cup D_{Re}$.

Definition 3 *A model, $\mathcal{M}$, is a structure*

$$\mathcal{M} = \langle W, S, D, \mathcal{B}, \mathcal{P}, \mathcal{I}, \pi, \Upsilon, \Omega, \Xi, \Phi, \Pi_1, \Pi_2 \rangle, where$$

- *W is a set of possible worlds;*
- *$S = \bigcup_{w \in W} S_w$ is the set of states occurred in W;*
- *D is the domain of individuals;*
- *$\mathcal{B} \subseteq W \times S \times D_{Ag} \times W$ is the belief accessibility relation;*
- *$\mathcal{P} \subseteq W \times S \times D_{Ag} \times 2^W$ assigns the frames of mind relation to each agent in D_{Ag} at different worlds and states. If $\mathcal{P}(w, s, A) = \{\mathcal{L}_1, \cdots, \mathcal{L}_k\}$, then each $\mathcal{L}_i (1 \leq i \leq k)$ contains the set of worlds that agent A, in that frame of mind, considers possible from state s of world w;*
- *$\mathcal{I} \subseteq W \times S \times D_{Ag} \times W$ is the intention accessibility relation. $\mathcal{I}(w, s, A)$ contains the set of worlds that agent A considers possible from state s of world w;*
- *π is an embedding dynamic graph for relating intentions. π is $\emptyset$ initially;*
- *$\Upsilon \subseteq W \times S \times D_{Ag} \times Pred \times D_{Re}$ establishes relations between recipes leading-to a proposition with agents at different worlds and states. $\Upsilon(w, s, A, p)$ is the set of recipes that agent A can use at state s of w to bring about p;*
- *$\Omega \subseteq D_{Ac} \times D_{Ag}$ relates act-types and agents. $\Omega(\alpha)$ gives the set of agents each can take the role of doing action α;*
- *Ξ interprets constants; and*
- *$\Phi \subseteq W \times S \times Pred \times D^k$ interprets predicates (Φ preserves arity);*
- *$\Pi_1 \subseteq D_A \times F_{state} \times F_{state}$ assigns initial intentions-that;*
- *$\Pi_2 \subseteq D_A \times AE \times F_{state}$ assigns initial intentions-to.*

Fagin & Halpern (1988) introduce the notion of "non-interacting clusters" of beliefs, where a belief held in one cluster or frame of mind may contradict a belief held in another cluster. Drawing upon this idea, we take potential intentions (the relation $\mathcal{P}$ above) as partitioning the possible futures into various "frames of mind".

The dynamic graph π captures the evolution of agent intentions. A dynamic theory of intention evolution can be very complicated. Here, we simply assume that π is populated with three kinds of state-indexed relations: upgrade relation ($\Cup$), originated-from relation ($\rightsquigarrow$), and elaboration relation ($\twoheadrightarrow$). Below we give some semantic rules for populating π with these relations.

Rule 1 *If at state s of world w, an agent A has a potential intention* $\mathsf{Pot.Int.Tx}(A, \alpha, C)$*, and there is no intention conflict resulting from this potential intention, then add* $\mathsf{Pot.Int.Tx}(A, \alpha, C) \Cup_{w,s} \mathsf{Int.Tx}(A, \alpha, C)$ *to π.*

By using Rule 1, the original potential intention is upgraded to a full-fledged intention.

The next rule states how potential intentions are derived from full-fledged intentions. Let $hasRole(A, \alpha) \triangleq \exists \varrho \in Role(\alpha) \cdot A \in \Omega(\varrho)$. Predicate $contains(e_1, e_2)$ is true iff e_2 is a sub action-expression of e_1; $doer(A, \alpha)$ represents that agent A is the doer of action α.

Rule 2 (deriving potential intentions)
(1) Add $\mathsf{Int.Th}(A, \phi, C) \rightsquigarrow_{w,s} \mathsf{Pot.Int.To}(A, \alpha, C')$ *to π where $C' = C \wedge \mathsf{Int.Th}(A, \phi, C)$, if at state s of world w agent A has the intention with respect to ϕ, and $\exists \gamma \cdot \gamma \in \Upsilon(w, s, A, \phi) \wedge \alpha = e_\gamma$;*
(2) Add $\mathsf{Int.To}(A, e, C) \rightsquigarrow_{w,s} \mathsf{Pot.Int.To}(A, \alpha, C')$ *to π where $C' = C \wedge \mathsf{Int.To}(A, e, C)$, if at state s of world w agent A has the intention to do e, and $\exists \beta \cdot contains(e, \alpha|\beta) \wedge hasRole(A, \alpha)$;*
(3) Add $\mathsf{Int.To}(G, e, C) \rightsquigarrow_{w,s} \mathsf{Pot.Int.To}(A, \alpha, C')$ *to π where $C' = C \wedge \mathsf{Int.To}(G, e, C) \wedge isDoer(A, \alpha)$, if at state s of world w, the group G has the intention to do e, and $contains(e, \alpha) \wedge \neg resolved(\alpha, G) \wedge (A \in G) \wedge hasRole(A, \alpha)$, where $resolved(\alpha, G)$ is defined in Fig. 2.*

Rule 2.(1) says that an agent adopts a potential intention to do α if α is the action expression of some recipe that can lead to ϕ. Rule 2.(2) says that an agent A adopts a potential intention to do α if A can take the role of α and α is a branch of some choice expression in e. Rule 2.(3) says that an agent A in a group G adopts a potential intention to do α if α is a sub-expression of e, A can take some role required in α, and the group G has not resolved who will take the role of doing α. Here, $isDoer(A, \alpha)$ in C' serves as an escape condition: A can drop the potential intention if later it turns out that the group designates another agent as the doer of α.

Elaboration relation is used to depict the means-ends structure of intentions. We first define a refinement relation.

Definition 4 *Given recipes $\gamma_1 = \langle e_{\gamma_1}, \rho_{\gamma_1} \rangle$ and $\gamma_2 = \langle e_{\gamma_2}, \rho_{\gamma_2} \rangle$, $refines(\gamma_2, \gamma_1, w, s)$ holds iff:*
(1) γ_1 and γ_2 lead to a same goal: $\exists \phi \exists A \cdot \{\gamma_1, \gamma_2\} \subseteq \Upsilon(w, s, A, \phi)$;
(2) Constraint ρ_{γ_2} is satisfiable given that ρ_{γ_1}: $\exists \mathcal{V}$ such that $\langle \mathcal{M}, \mathcal{V}, w, s \rangle, \rho_{\gamma_1} \models \rho_{\gamma_2}$ (cf. Fig. 2 for the def. of $\models$); and

(3) e_{γ_2} reifies e_{γ_1}: $e_{\gamma_1} \rhd^+ e_{\gamma_2}$, where e_{γ_2} results from e_{γ_1} by applying $\rhd$ one or more times, where $\rhd$ is defined as:
• if predicate $contains(e, e_1|e_2)$ holds, then $e \rhd e[e_1|e_2, e_1]$ and $e \rhd e[e_1|e_2, e_2]$, where $e[e_1|e_2, e_1]$ is e with the occurrence of '$e_1|e_2$' replaced by e_1;
• if $contains(e, e_1^*)$, then $e \rhd e[e_1^*, e_1^k]$, where e_1^k is a sequence of e_1 with a fixed length k;
• if $contains(e, \gamma)$ and $refines(\gamma', \gamma, w, s)$, then $e \rhd e[\gamma, \gamma']$;
• if $contains(e, \phi!)$ and $\exists A \exists \gamma \cdot \gamma \in \Upsilon(w, s, A, \phi)$, then $e \rhd e[\phi!, \gamma]$.

The basic idea behind the notion of recipe refinement is that rational agents tend to adopt concrete intentions to pursue general ones.

Rule 3 *Add* $\mathsf{Int.To}(A, e_{\gamma_1}, C) \twoheadrightarrow_{w,s} \mathsf{Int.To}(A, e_{\gamma_2}, C')$ *to π where $C' = C \wedge refines(\gamma_2, \gamma_1, w, s)$, if at state s of world w agent A has the intention to do e_{γ_1}, and $\exists \gamma_2 \cdot refines(\gamma_2, \gamma_1, w, s)$.*

Rule 3 says that an agent adopts an intention to do e_{γ_2} if the recipe γ_2 refines γ_1–the one being followed by the agent. Here, $refines(\gamma_2, \gamma_1, w, s)$ in C' accounts for the adoption of the more elaborated intention.

More rules, which are omitted here, can be given to further extend $\twoheadrightarrow$ to accommodate the "elaboration" relation relative to decomposition and specialization as considered by Konolige & Pollack (1993).

We now define the notion of 'sub-world'.

Definition 5 *A world w' is a sub-world of the world w, denoted by $w' \prec w$, if and only if*
(a) $S_{w'} \subseteq S_w$; (b) $R_{w'} \subseteq R_w$;
(c) $\forall s \in S_{w'}, \Phi(w', s) = \Phi(w, s)$;
(d) $\forall s \in S_{w'}, \forall A \in D_{Ag}, \forall v \in W, (w', s, A, v) \in \mathcal{B}$ iff $(w, s, A, v) \in \mathcal{B}$;
(e) $\forall s \in S_{w'}, \forall A \in D_{Ag}, \forall \mathcal{L} \in 2^W, (w', s, A, \mathcal{L}) \in \mathcal{P}$ iff $(w, s, A, \mathcal{L}) \in \mathcal{P}$;
(f) $\forall s \in S_{w'}, \forall A \in D_{Ag}, \forall v \in W, (w', s, A, v) \in \mathcal{I}$ iff $(w, s, A, v) \in \mathcal{I}$; and
(g) $\forall s \in S_{w'}, \forall A \in D_A, \forall p \in Pred, \forall \gamma \in D_{Re}, (w', s, A, p, \gamma) \in \Upsilon$ iff $(w, s, A, p, \gamma) \in \Upsilon$.

The Semantics

Notations: $w^r_{s_i}$ denotes the path in w starting from state s_i and defined by r—a linear sub-relation of R_w. $s_i <_r s_j$ denotes that s_j is the next state of s_i along r; $s_i <^*_r s_j$ denotes that s_j is a state accessible from state s_i along r. $[s_0, s_k)_r = \{s | s_0 <^*_r s <^*_r s_k \; and \; s \neq s_k\}$. We also use $s_i <^*_r s_j$ to refer to the segment of path r between s_i and s_j.

Let $act(s_i <^*_r s_j)$ be the sequence of primitive actions occurred in the path segment $s_i <^*_r s_j$. $run(\kappa, \alpha)$ holds if the action sequence κ is a run of action expression α.

The function $[\![\cdot]\!]$ gives the denotation of a term relative to Ξ and a sort-preserving variable assignment function $\mathcal{V}$: $[\![\tau]\!]$ is $\Xi(\tau)$ if $\tau \in Const$, and $\mathcal{V}(\tau)$ otherwise. Let $\mathcal{V}^x_c$ be an assignment function agreeing with $\mathcal{V}$ except for variable x: $\mathcal{V}^x_c(x) = c$, and $\forall y \neq x \cdot \mathcal{V}^x_c(y) = \mathcal{V}(y)$.

Satisfaction of formulas, denoted by $\models$, is given with respect to a model $\mathcal{M}$, a variable assignment $\mathcal{V}$, a world w, and a state s or a path w^r_s. Figure 2 gives the semantics of L. The semantics of first-order connectives, temporal operators, and Bel is straightforward. We here focus on the semantics of intentional operators.

Int.Th: An agent A intends that ϕ hold under context C, iff (i) ϕ is an ascribed intention ($(\phi, C) \in \Pi_1(A)$) and C holds; or (ii) $\mathsf{Int.Th}(A, \phi, C)$ is upgraded from, or connected with, some other (potential) intention (recorded in π), and for any v accessible from (w, s), (a) ϕ and C hold at the first state of v, (b) v is a sub-world of w, and (c) there exists a path from s to v_0. This semantics of intention differs from the existing approaches in three aspects. First, it leverages the features of both the representational and accessibility-relation approaches. Consequently, the model accommodates both ascribed intentions and dynamically generated intentions. Second, satisfying the intention-accessibility relation is no longer the sufficient condition for an agent to hold a non-ascribed intention. The formula itself must have some relation with other intention as recorded in π. This, again, is a strong representationalist feature. Third, to have a non-ascribed intention, any world accessible from (w, s) must be a sub-world of w and the world can be reached from s along some path in w. This establishes a reasonable connection between the accessibility relation and the temporal precedence relation. Thus, intuitively, if $\mathsf{Int.Th}(A, \phi, C)$ (where $(\phi, C) \notin \Pi_1(A)$) holds at (w, s), it must be the case that every intention-accessible world can be reached from state s after some effort.

An intention is dropped when the goal is satisfied or it becomes unachievable. The contexts of intentions can be used for such purposes. For example, add $\neg\mathsf{Bel}(A, \phi)$ to C when an agent A has an intention $\mathsf{Int.Th}(A, \phi, C)$ where $(\phi, C) \in \Pi_1(A)$. Then the intention can be dropped when $\mathsf{Bel}(A, \phi)$ holds (e.g., remove ϕ from Π_1). Similarly, the context of an intention could contain the source intention; the intention is abandoned when the source is dropped. The detail of this topic is out of the scope of this paper.

Int.To: An agent A intends to do α under context C, iff (i) α is an ascribed intention ($(\alpha, C) \in \Pi_2(A)$) and C holds; or (ii) $\mathsf{Int.To}(A, \alpha, C)$ is upgraded from, or connected with, some other (potential) intention (recorded in π), and for any v accessible from (w, s), (a) $Done(\alpha)$ and C hold at the first state of v, (b) v is a sub-world of w, and (c) there exists a path from s to v_0. Here, $Done(\alpha)$ is defined backward from s along a path: there exists a past state s' such that the action sequence occurred in the path segment $s' <^*_r s$ is a run of α.

Pot.Int.Th: An agent potentially intends that ϕ hold under context C, iff (a) the agent does not have an intention regarding ϕ yet; (b) $\mathsf{Pot.Int.Th}(A, \phi, C)$ is originated from some other intention (recorded in π), and there exists a frame $\mathcal{L}_i$ of mind accessible from (w, s) such that for any $v \in \mathcal{L}_i$: (c) ϕ and C hold at the first state of v; (d) v is a sub-world of w; and (e) there exists a path from s to v_0. Intuitively, if $\mathsf{Pot.Int.Th}(A, \phi, C)$ holds at (w, s), it must be the case that there exists a frame of A's mind accessible from (w, s) such that every world in the frame can be reached after some effort (along some path) from state s in world w. Such defined, the model allows inconsistent potential intentions: they can hold in different frames of mind.

Pot.Int.To: An agent potentially intends to do α under

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \mathsf{true} \quad (1)$$

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models p(\tau_1, \cdots, \tau_n) \textit{ iff } \langle [\![\tau_1]\!], \cdots, [\![\tau_n]\!]\rangle \in \Phi(w, s, p) \quad (2)$$

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \neg\phi \textit{ iff } \langle \mathcal{M}, \mathcal{V}, w, s\rangle \not\models \phi \quad (3)$$

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \phi \vee \psi \textit{ iff } \langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \phi \textit{ or } \langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \psi \quad (4)$$

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \exists y \cdot \phi \textit{ iff } \exists m \in D \textit{ such that } \langle \mathcal{M}, \mathcal{V}^y_m, w, s\rangle \models \phi \quad (5)$$

$$\langle \mathcal{M}, \mathcal{V}, w^r_s\rangle \models \phi \textit{ iff } \langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \phi, \textit{where } \phi \textit{ is a state formula} \quad (6)$$

$$\langle \mathcal{M}, \mathcal{V}, w^r_s\rangle \models \mathsf{X}\phi \textit{ iff } \langle \mathcal{M}, \mathcal{V}, w^r_{s'}\rangle \models \phi, \textit{where } s <_r s' \quad (7)$$

$$\langle \mathcal{M}, \mathcal{V}, w^r_s\rangle \models \phi \mathsf{U} \psi \textit{ iff } (a)\ \exists s_k \cdot s <^*_r s_k \textit{ such that } \langle \mathcal{M}, \mathcal{V}, w^r_{s_k}\rangle \models \psi, \textit{ and } \forall s' \in [s, s_k)_r, \langle \mathcal{M}, \mathcal{V}, w^r_{s'}\rangle \models \phi, \textit{or}$$
$$(b)\ \forall s' \cdot s <^*_r s', \langle \mathcal{M}, \mathcal{V}, w^r_{s'}\rangle \models \phi \quad (8)$$

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \mathsf{E}\phi \textit{ iff there exists a path } w^r_s \textit{ such that } \langle \mathcal{M}, \mathcal{V}, w^r_s\rangle \models \phi \quad (9)$$

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models (x \in G) \textit{ iff } [\![x]\!] \in [\![G]\!] \quad (10)$$

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models (\alpha \in \gamma) \textit{ iff } \textit{contains}([\![e_\gamma]\!], [\![\alpha]\!]) \quad (11)$$

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models (\gamma_1 \sqsupseteq \gamma_2) \textit{ iff } \textit{refines}([\![\gamma_1]\!], [\![\gamma_2]\!], w, s) \quad (12)$$

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \mathsf{Bel}(A, \phi) \textit{ iff } \forall v \in \mathcal{B}(w, s, A) \cdot \langle \mathcal{M}, \mathcal{V}, v, v_0\rangle \models \phi \quad (13)$$

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \mathsf{Int.Th}(A, \phi, C) \textit{ iff } (i)\ (\phi, C) \in \Pi_1(A) \textit{ and } \langle \mathcal{M}, \mathcal{V}, w, s\rangle \models C, \textit{ or}$$
$$(ii)\ [\exists \eta, w', s' \cdot (\eta \uplus_{w',s'} \mathsf{Int.Th}(A, \phi, C) \in \pi) \vee (\eta \twoheadrightarrow_{w',s'} \mathsf{Int.Th}(A, \phi, C) \in \pi)], \textit{and}$$
$$\forall v \in \mathcal{I}(w, s, A),\ (a)\ \langle \mathcal{M}, \mathcal{V}, v, v_0\rangle \models \phi \wedge C, (b)\ v \prec w, \textit{ and } (c)\ \exists r \cdot s <^*_r v_0 \quad (14)$$

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \mathsf{Int.To}(A, \alpha, C) \textit{ iff } (i)\ (\alpha, C) \in \Pi_2(A) \textit{ and } \langle \mathcal{M}, \mathcal{V}, w, s\rangle \models C, \textit{ or}$$
$$(ii)\ [\exists \eta, w', s' \cdot (\eta \uplus_{w',s'} \mathsf{Int.To}(A, \alpha, C) \in \pi) \vee (\eta \twoheadrightarrow_{w',s'} \mathsf{Int.To}(A, \alpha, C) \in \pi)], \textit{and}$$
$$\forall v \in \mathcal{I}(w, s, A),\ (a)\ \langle \mathcal{M}, \mathcal{V}, v, v_0\rangle \models \textit{Done}(\alpha) \wedge C, (b)\ v \prec w, \textit{ and } (c)\ \exists r \cdot s <^*_r v_0 \quad (15)$$

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \mathsf{Pot.Int.Th}(A, \phi, C) \textit{ iff } (a)\ \langle \mathcal{M}, \mathcal{V}, w, s\rangle \not\models \mathsf{Int.Th}(A, \phi, C), (b)\ \exists \eta, w', s' \cdot (\eta \rightsquigarrow_{w',s'} \mathsf{Pot.Int.Th}(A, \phi, C) \in \pi),$$
$$\textit{and } \exists \mathcal{L}_i \in \mathcal{P}(w, s, A), \forall v \in \mathcal{L}_i,\ (c)\ \langle \mathcal{M}, \mathcal{V}, v, v_0\rangle \models \phi \wedge C, (d)\ v \prec w, \textit{ and } (e)\ \exists r \cdot s <^*_r v_0 \quad (16)$$

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \mathsf{Pot.Int.To}(A, \alpha, C) \textit{ iff } (a)\ \langle \mathcal{M}, \mathcal{V}, w, s\rangle \not\models \mathsf{Int.To}(A, \alpha, C), (b)\ \exists \eta, w', s' \cdot (\eta \rightsquigarrow_{w',s'} \mathsf{Pot.Int.To}(A, \alpha, C) \in \pi),$$
$$\textit{and } \exists \mathcal{L}_i \in \mathcal{P}(w, s, A), \forall v \in \mathcal{L}_i,\ (c)\ \langle \mathcal{M}, \mathcal{V}, v, v_0\rangle \models \textit{Done}(\alpha) \wedge C, (d)\ v \prec w, \textit{ and } (e)\ \exists r \cdot s <^*_r v_0 \quad (17)$$

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \mathsf{IntX}(G, \alpha, C) \textit{ iff } \forall x \in G, \langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \mathsf{IntX}(x, \alpha, C \wedge \mathsf{IntX}(G, \alpha, C)),$$
$$\textit{where } \mathsf{IntX} \in \{\mathsf{Int.To}, \mathsf{Int.Th}, \mathsf{Pot.Int.To}, \mathsf{Pot.Int.Th}\} \quad (18)$$

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \textit{Done}(\alpha) \textit{ iff } \exists r, s' \cdot s' <^*_r s, \textit{and } \textit{run}(\textit{act}(s' <^*_r s), \alpha) \quad (19)$$

$$\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \textit{resolved}(\alpha, G) \textit{ iff } \forall \varrho \in \textit{Role}(\alpha), \exists A \in G \cap \Omega(\varrho) \textit{ such that } \forall x \in G \cdot \langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \mathsf{Bel}(x, \textit{isDoer}(A, \varrho)) \quad (20)$$

Figure 2: The formal semantics

context C, iff (a) the agent does not have an intention to do α yet; (b) $\mathsf{Pot.Int.To}(A, \alpha, C)$ is originated from some other intention (recorded in π), and there exists a frame $\mathcal{L}_i$ of A's mind accessible from (w, s) such that for any $v \in \mathcal{L}_i$: (c) $\textit{Done}(\alpha)$ and C hold at the first state of v; (d) v is a sub-world of w; and (e) there exists a path from s to v_0.

Group intentions: A group G has an intention regarding α iff each agent in G has an intention regarding α under the context that the individual intention originates from, and is governed by, the group intention. An intention ultimately leads to the performing of actions. Since group intentions entail individual intentions, the fulfillment of $\mathsf{Int.To}(G, e, C)$ depends on the fulfillment of $\mathsf{Int.To}(A, e, C')$ for each $A \in G$. Without going into detail, we here assume that by fulfilling $\mathsf{Int.To}(A, e, C')$, agent A simply performs those primitive actions in e that are within A's capability. Such assumed, if $\forall \varrho \in \textit{Role}(e) \cdot A \in \Omega(\varrho) \vee B \in \Omega(\varrho)$, then $\mathsf{Int.To}(A, e, C')$ and $\mathsf{Int.To}(B, e, C')$ jointly will get e done.

Relationship among intentions

The above model allows us to spell out the relations between intentions and potential intentions.

Definition 6 *The structure $\mathcal{M}$ is upgrade-allowable iff: if $\mathsf{Pot.Int.Tx}(A, \alpha, C) \uplus_{w,s} \mathsf{Int.Tx}(A, \alpha, C) \in \pi$, then (1) $\langle \mathcal{M}, \mathcal{V}, w, s^{-1}\rangle \models \mathsf{Pot.Int.Tx}(A, \alpha, C)$, where $s^{-1} R_w s$; (2) $\langle \mathcal{M}, \mathcal{V}, w, s\rangle \models \mathsf{Int.Tx}(A, \alpha, C)$; and (3) $\exists F \subseteq \mathcal{P}(w, s^{-1}, A)$ such that F is consistent with respect to α, and $\mathcal{I}(w, s, A) \subseteq \bigcup_{\mathcal{L}_i \in F} \mathcal{L}_i$.*

Definition 7 *The structure $\mathcal{M}$ is proactive-allowable iff: (1) if $\mathsf{Int.Th}(A, \phi, C) \rightsquigarrow_{w,s} \mathsf{Pot.Int.To}(A, \alpha, C')$, then $\exists \mathcal{L}_i \in \mathcal{P}(w, s, A)$ such that $\mathcal{L}_i \subseteq \mathcal{I}(w, s, A)$ and $\langle \mathcal{M}, \mathcal{V}, w, v\rangle \models \phi \wedge \textit{Done}(\alpha)$ for all $v \in \mathcal{L}_i$; (2) if $\mathsf{Int.To}(A, e, C) \rightsquigarrow_{w,s} \mathsf{Pot.Int.To}(A, \alpha, C')$, then (a) $\exists \mathcal{L}_i \in \mathcal{P}(w, s, A)$ such that $\mathcal{L}_i \subseteq \mathcal{I}(w, s, A)$ and $\langle \mathcal{M}, \mathcal{V}, w, v\rangle \models \textit{Done}(e) \wedge \textit{Done}(\alpha)$ for all $v \in \mathcal{L}_i$; and (b) $\exists \mathcal{L}_j \in \mathcal{P}(w, s, A)$ such that $\mathcal{L}_j \subseteq \mathcal{I}(w, s, A)$ and $\langle \mathcal{M}, \mathcal{V}, w, v\rangle \models \textit{Done}(e) \wedge \neg\textit{Done}(\alpha)$ for all $v \in \mathcal{L}_j$; (3) if $\mathsf{Int.To}(G, e, C) \rightsquigarrow_{w,s} \mathsf{Pot.Int.To}(A, \alpha, C')$, then $\exists \mathcal{L}_i \in \mathcal{P}(w, s, A)$ such that $\mathcal{L}_i \subseteq \mathcal{I}(w, s, A)$ and for all $v \in \mathcal{L}_i$, $\langle \mathcal{M}, \mathcal{V}, w, v\rangle \models \textit{Done}(e) \wedge \textit{Done}(\alpha)$ and $\exists \varrho \in \textit{Role}(\alpha)$ such that $\varrho \in \textit{act}(s <^* v)$ and $\textit{isDoer}(A, \varrho)$.*

Definition 8 *The structure $\mathcal{M}$ is refinement-allowable iff:*

if $\mathsf{Int.To}(A, e_{\gamma_1}, C_1) \twoheadrightarrow_{w,s} \mathsf{Int.To}(A, e_{\gamma_2}, C_2)$*, then* $\exists \phi \cdot \gamma_1 \in \Upsilon(w, s, A, \phi)$ *such that* $\forall v \in \mathcal{I}(w, s, A)$, $\langle \mathcal{M}, \mathcal{V}, w, v \rangle \models \phi \wedge Done(e_{\gamma_2})$.

Let $\mathcal{M}_0$ be a model that is upgrade-allowable, proactive-allowable, and refinement allowable. We examine the relations between potential intentions and intentions in $\mathcal{M}_0$.

Proposition 1

1. $\mathcal{M}_0$ allows agents to hold conflicting potential intentions;

2. An intention cannot coexist with a potential intention towards the same ends:

$\langle \mathcal{M}_0, \mathcal{V}, w, s \rangle \not\models \mathsf{Int.Tx}(A, \alpha, C) \wedge \mathsf{Pot.Int.Tx}(A, \alpha, C)$*;*

3. Agents proactively choose actions relative to their competence:

$\langle \mathcal{M}_0, \mathcal{V}, w, s \rangle \models \mathsf{Int.To}(A, \alpha|\beta, C) \wedge hasRole(A, \alpha) \rightarrow \mathsf{Pot.Int.To}(A, \alpha, C \wedge \mathsf{Int.To}(A, \alpha|\beta, C))$*;*

$\langle \mathcal{M}_0, \mathcal{V}, w, s \rangle \models \mathsf{Int.To}(G, \alpha; \beta, C) \wedge (A \in G) \wedge hasRole(A, \alpha) \wedge \neg resolved(\alpha, G) \rightarrow \mathsf{Pot.Int.To}(A, \alpha, C \wedge \mathsf{Int.To}(A, \alpha|\beta, C))$.

We then consider the relations between Bel and Int.Tx.

Definition 9 *Define the following constraints on $\mathcal{B}$ and $\mathcal{I}$:*

(R_1) $\forall w \forall s \forall A \exists v$, $v \in \mathcal{I}(w, s, A) \cap \mathcal{B}(w, s, A)$*;*

(R_2) $\forall w \forall s \forall A \forall v \in \mathcal{B}(w, s, A)$, $\mathcal{I}(w, s, A) \subseteq \mathcal{I}(w, v, A)$*;*

(R_3) $\forall w \forall s \forall A \forall v \in \mathcal{I}(w, s, A)$, $\mathcal{B}(w, v, A) \subseteq \mathcal{I}(w, s, A)$.

(R_1) is actually the weak-realism constraint (Rao & Georgeff 1995). Let $\mathcal{M}_{\{i\}}$ be $\mathcal{M}_0$ satisfying (R_i). We have the following properties about belief-intention relations.

Proposition 2

$\langle \mathcal{M}_{\{1\}}, \mathcal{V}, w, s \rangle \models \mathsf{Int.Th}(A, \phi, C) \rightarrow \neg \mathsf{Bel}(A, \neg\phi)$*;*

$\langle \mathcal{M}_{\{1\}}, \mathcal{V}, w, s \rangle \models \mathsf{Int.To}(A, \alpha, C) \rightarrow \neg \mathsf{Bel}(A, \neg Done(\alpha))$*;*

$\langle \mathcal{M}_{\{2\}}, \mathcal{V}, w, s \rangle \models \mathsf{Bel}(A, \mathsf{Int.Th}(A, \phi, C)) \rightarrow \mathsf{Int.Th}(A, \phi, C)$*;*

$\langle \mathcal{M}_{\{3\}}, \mathcal{V}, w, s \rangle \models \mathsf{Int.Th}(A, \phi, C) \rightarrow \mathsf{Int.Th}(A, \mathsf{Bel}(A, \phi), C)$*;*

$\langle \mathcal{M}_{\{2,3\}}, \mathcal{V}, w, s \rangle \models \mathsf{Bel}(A, \mathsf{Int.Th}(A, \phi, C)) \rightarrow \mathsf{Int.Th}(A, \mathsf{Bel}(A, \phi), C)$.

In Proposition 2, the first two say that in $\mathcal{M}_{\{1\}}$, an agent does not intend what it believes impossible (This is the Axiom 1 of the SharedPlans theory (Grosz & Kraus 1999)). The third one states that an agent does have the intentions if it believes it so intends (This is the Axiom (A3) of the SharedPlans theory (Grosz & Kraus 1996)). We can give constraints similar to (R_2) to validate the Axioms (i.e., A2 and A4 (Grosz & Kraus 1996)) that relate Bel with Int.To and potential intentions. The fourth one states that an agent cannot intend ϕ without intending to believe ϕ (Sadek 1992). The (R_2) and (R_3) together validate the fifth one, which shows certain commutativity between Int.Th and Bel.

Summary

While Int.To and Int.Th play the functional roles of intentions (Bratman 1990), Pot.Int.To and Pot.Int.Th only represent potential commitments. In this paper, we give a formal semantics to the four intentional operators of the SharedPlans theory. Our model considers the dynamic relationship among the four types of intention attitudes, drawing upon both the representationalist approach and the accessibility-based approach. Using the formal semantics defined in this paper, one can validate many of the proposed properties of the SharedPlans theory and other belief-intention driven multi-agent collaboration teamworks.

References

Bratman, M. E. 1987. *Intention, Plans, and Practical Reason*. Harvard University Press.

Bratman, M. E. 1990. What is intention? In *Intentions in Communication*, 15–31. MIT Press.

Cohen, P. R., and Levesque, H. J. 1990. Intention is choice with commitment. *Artificial Intelligence* 42:213–261.

Emerson, E. A. 1990. Temporal and modal logic. In *Handbook of theoretical computer science (vol. B): formal models and semantics*. MIT Press. 995–1072.

Fagin, R., and Halpern, J. Y. 1988. Beliefs, awareness and limited reasoning. *Artificial Intelligence* 34:39–76.

Grosz, B., and Kraus, S. 1996. Collaborative plans for complex group actions. *Artificial Intelligence* 86:269–358.

Grosz, B., and Kraus, S. 1999. The evolution of shared-plans. In Rao, A., and Wooldridge, M., eds., *Foundations and Theories of Rational Agencies*, 227–262.

Grosz, B., and Sidner, C. 1990. Plans for discourse. In Cohen, P.; Morgan, J.; and Pollack, M., eds., *Intentions in communication*. MIT Press. 417–444.

Halpern, J. Y., and Moses, Y. 1992. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence* 54(3):319–379.

Herzig, A., and Longin, D. 2002. A logic of intention with cooperation principles and with assertive speech acts as communication primitives. In *AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems*, 920–927.

Konolige, K., and Pollack, M. E. 1993. A representationalist theory of intention. In Bajcsy, R., ed., *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI-93)*, 390–395. Chambéry, France: Morgan Kaufmann publishers Inc.: San Mateo, CA, USA.

Rao, A., and Georgeff, M. 1995. Formal models and decision procedures for multi-agent systems. Technical Report 61, Australian Artificial Intelligence Institute, Melbourne, Australia.

Sadek, M. 1992. A study in the logic of intention. In Nebel, B.; Rich, C.; and Swartout, W., eds., *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, 462–473. Morgan Kaufmann publishers Inc.: San Mateo, CA.

Searle, J. R. 1983. *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.

Singh, M. P., and Asher, N. M. 1993. A logic of intentions and beliefs. *Journal of Philosophical Logic* 22:513–544.

Singh, M. P. 1993. Intentions for multiagent systems. Technical Report KBNL-086-93, MCC, Information Systems Division.

Wooldridge, M., and Jennings, N. R. 1994. Towards a theory of cooperative problem solving. In *Proc. Modelling Autonomous Agents in a Multi-Agent World (MAAMAW-94)*, 15–26.