Delegation and mental states*

Emiliano Lorini
ISTC-CNR, Rome (Italy)
emiliano.lorini@istc.cnr.it

Nicolas Troquard
IRIT, Toulouse (France) and
LOA-ISTC-CNR, Trento (Italy)
troquard@irit.fr

Andreas Herzig
IRIT-CNRS, Toulouse (France)
herzig@irit.fr

Cristiano Castelfranchi
ISTC-CNR, Rome (Italy)
c.castelfranchi@istc.cnr.it

ABSTRACT

In the recent literature on multiagent systems there have been several proposals of formal systems for reasoning about delegation. Most of these approaches have dealt with the concept of delegation leaving mental states such as beliefs, goals and intentions out of consideration. The aim of this paper is to develop a formal approach for reasoning about delegation by modeling intentions and beliefs of the delegating agent in an explicit way. We present a logic where it is possible to investigate the relations between the concept of *Intention to be* and the concept of *Delegation*.

Categories and Subject Descriptors

I.2.11 [**Distributed artificial intelligence**]: multiagent systems

General Terms

Theory

Keywords

intention, delegation, BDI, logics for agent systems, modal logic

1. INTRODUCTION

The 1990 formalization of Bratman's theory of intention [2] by Cohen and Levesque (C&L) [3] was designated one of the most influential papers in the domain at AAMAS'06. Their approach is based on a logical framework integrating the concepts of belief, action, time, and preference. In that framework they successively define several notions of goal, and finally intention. An agent i has the **goal** that ϕ if ϕ holds in the future of all of i's preferred histories ($Pref_i F\phi$), where the formal substratum for histories is Linear Temporal Logic LTL.[1] An **achievement goal** of agent i is a goal of i of which i believes it is not achieved yet:

$$AGoal_i^{CL}\phi \stackrel{\text{def}}{=} Pref_i F\phi \wedge Bel_i\neg\phi.$$

Then, a **persistent goal** is an achievement goal that persists until it is either achieved, or believed to be impossible:

$$PGoal_i^{CL}\phi \stackrel{\text{def}}{=} AGoal_i^{CL}\phi \wedge (Bel_i\phi \vee Bel_i G\phi) Before \neg Pref_i F\phi.$$

Finally, C&L define an **intention that** ϕ as a persistent goal to the achievement of which the agent actively contributes, in the sense that in every preferred history there must be some action α whose author is i and which brings about ϕ. Noting $i{:}\alpha$ such an action and using quantification over actions this can be written:

$$Int_i^{CL}\phi \stackrel{\text{def}}{=} PGoal_i\phi \wedge Pref_i F\exists i{:}\alpha\langle i{:}\alpha\rangle\phi,$$

where $\langle i{:}\alpha\rangle\phi$ reads "i does action α and ϕ holds after α's occurrence". While being too strong in some respect[2], C&L's definition is *too weak* because it lacks a causal connection between the action and the goal. Indeed, suppose that agent i wants to go to a shoe shop on Saturday: $Pref_i F\langle i{:}GoToShoeShop\rangle\top$. Moreover suppose that i has the persistent goal that for all weekend it will be sunny: $PGoal_i^{CL} Sunny$. The previous preference and the persistent goal are completely unrelated, that is, *going to a shoe shop on Saturday* is not part of a plan for achieving the result *for all weekend it will be sunny*. Agent i is simply endorsing two different goals at the same time. From the fact that i wants to perform the action of going to a shoe shop on Saturday and the fact that he has the persistent goal that for all weekend it will be sunny it follows that each of i's preferred histories has the action *going to a shoe shop* of i leading to a state where the fact *Sunny* holds: $Pref_i F\langle i{:}GoToShoeShop\rangle Sunny$. From the previous it follows that each of i's preferred histories has some action of i leading to a state where the fact *sunny* holds: $Pref_i F\exists i{:}\alpha\langle i{:}\alpha\rangle Sunny$. According to Cohen and Levesque's definition of *intention to be*, i has the intention that *it will be sunny*: $Int_i^{CL} Sunny$. This consequence seems to us unacceptable. There is common agreement in philosophy that we cannot reasonably say that we intend that some event occurs when we believe that the occurrence of this event is independent of us. According to Searle for instance I cannot

*This work has been done with the support of the Interdisciplinary Laboratory on Interacting Knowledge Systems (ILIKS).

AAMAS'07, May 14–18, 2007, Honolulu, Hawai'i, USA.

[1] $G\phi$ reads "ϕ holds henceforth", $F\phi$ reads "ϕ will eventually hold", $X\phi$ reads "ϕ will hold at the next time point", and $\phi Before \psi$ reads "ϕ holds before ψ holds", $Pref_i\phi$ reads "i wants ϕ to be true".

[2] See [7] for a criticism.

978-81-904262-7-5 (RPS)

say that I intend that it will rain or I intend that the sun will rise etc... [8].

The causal connection between action and goal which is missed in C&L's logic is exactly what theories of agency such as Belnap, Horty, Chellas et col.'s 'seeing-to-it-that' STIT and Kanger, Pörn et col.'s 'bringing-it-about' provides.

Our aim in this paper is to combine C&L's approach with Chellas' STIT operator, and argue that the resulting logic is rich enough to provide a satisfactory account of the notion of *intention to be* and of *delegation*.

2. A LOGIC OF AGENCY AND MENTAL STATES

We have a standard possible worlds semantics for our framework, where each modal operator has logic K and is thus a normal modality. It is essentially Chellas' STIT logic of agency [1]. Models of intention are tuples:

$$\langle Mom, <, ATM, AGT, Choice, B, P, v\rangle$$

$\langle Mom, <\rangle$ is a *branching-time*, *discrete* structure. A *history* is a maximal $<$-ordered subset of Mom. $Hist$ is the set of all histories, and for $w \in Mom$, H_w is the set of histories passing through w. An index (alias context) is a moment-history pair w/h. ATM is a set of atoms, which are evaluated w.r.t. a context: $v : ATM \longrightarrow 2^{Mom \times Hist}$. AGT is the finite set of all agents. $Choice : 2^{AGT} \times Mom \longrightarrow 2^{2^{Hist}}$ is a function mapping each set of agents and moment w into a *partition* of H_w. The equivalence classes belonging to $Choice^w_J$ can be thought of as possible choices available to agents $J \in AGT$ at moment w. For $h \in H_w$, $Choice^w_J(h)$ is the particular choice of J at context w/h. We have:
$M, w/h \models Stit_J\phi$ iff $M, w/h' \models \phi$ for every $h' \in Choice^w_J(h)$. Every $Stit_J$ is thus an S5 modality. It reads "agents J see to it that ϕ". The choice functions are constraints s.t. $Choice^w_J(h) = \cap_{i\in J} Choice^w_i(h)$, and agents' choices are independent, hence, given an arbitrary selection function s s.t. $s(w, i) \in Choice^w_i$, $\cap_{a\in AGT} s(w, i) \neq \emptyset$.

STIT moreover has an operator of historic necessity $\Box$ and its dual $\Diamond$, whose accessibility relation is reflexive, transitive and euclidian (S5). We have:

$$M, w/h \models \Box\phi \text{ iff } M, w/h' \models \phi \text{ for all } h' \in H_w.$$

B_i and P_i are accessibility relations between contexts. They are serial, transitive and euclidian for every agent i. We build upon them the KD45 necessity operators Bel_i and $Pref_i$, defined as usual. We impose that B_i contains P_i, but also introspection: if wB_iw' then $P_i(w) = P_i(w')$. The temporal operators X and G are from LTL. $M, m/h \models X\phi$ iff $M, w'/h \models \phi$, w' being the immediate successor of w in history h. The accessibility relation for X is functional and serial, and the one for G is the reflexive and transitive closure for that of X. $F\phi$ abbreviates $\neg G\neg\phi$.
The following formulae are valid:
(Stit) S5 axioms for $Stit_J$;
(BoxStit) $\Box\phi \rightarrow Stit_i\phi$;
(Monotony) $Stit_I\phi \rightarrow Stit_J\phi$, for $I \subseteq J$;
(LTL) axioms of LTL (see [5]);
(Bel/Pref) KD45 axioms for Bel_i and $Pref_i$;
(Inclusion) $Bel_i\phi \rightarrow Pref_i\phi$;
(Pos. introspection) $Pref_i\phi \rightarrow Bel_iPref_i\phi$;
(Neg. introspection) $\neg Pref_i\phi \rightarrow Bel_i\neg Pref_i\phi$.

3. INTENTION TO BE

We start with a definition of **achievement goal** similar to C&L's definition. An **achievement goal** of agent i is a goal of which i does not believe it is already achieved:[3]

$$AGoal_i\phi \stackrel{\text{def}}{=} Pref_iF\phi \wedge \neg Bel_i\phi.$$

Our definition of **intention to be** is:

$$Int_i\phi \stackrel{\text{def}}{=} AGoal_i\phi \wedge Bel_i\neg Stit_{AGT\setminus\{i\}}F\phi.$$

Therefore according to the previous definition an agent i intends that ϕ iff agent i has the achievement goal that ϕ and believes that his intervention is needed in order to produce ϕ. $Bel_i\neg Stit_{AGT\setminus\{i\}}F\phi$ is called *dependence belief*.

According to our definition of *intention to be*, an agent i cannot have the intention that it will rain or the intention that the sun will rise and so on. Indeed events such as *it rains*, *the sun rises* etc. are events ϕ that satisfy the following property of independence from an arbitrary agent i:

$$Indep(\phi, i) \stackrel{\text{def}}{=} F\phi \rightarrow Stit_{AGT\setminus\{i\}}F\phi.$$

This means that events such as *it rains*, *the sun rises* are events whose possible future occurrence does not depend on agent i's behavior. For instance, if it is the case that *the sun rises* then this fact is true independently from what agent i does: $Indep(SunRises, i)$. Now given an event ϕ (such as *the sun rises* or *it rains*) that an agent (reasonably) believes to be independent of himself, can we say that the agent intends that ϕ? According to our definition this is not possible. Indeed the formula $Bel_iIndep(\phi, i) \wedge Int_i\phi \rightarrow \bot$ is valid in our logic. It is in this sense that we improve over C&L.

In our view the crucial aspect of the notion of *intention to be* is the fact that this is inseparable from means-end reasoning and deliberation. In order to understand what *an agent intends that* ϕ means, we must focus on the agent's planning activity for the achievement of ϕ. Our claim is the following: an agent i intends that ϕ only if he has decided to pursue some plans for achieving ϕ (viz. he intends to do something in order to achieve ϕ) or at least he is convinced that he must do something in order to achieve ϕ. We claim that this is the crucial aspect of the notion of *intention to be* and that it is nicely expressed by the *dependence belief*. Therefore if i intends that ϕ then either he has already decided to pursue a specific plan in order to achieve ϕ (viz. i intends to do something in order to achieve ϕ), or he is starting to build a plan in order to achieve ϕ.[4] Our notion of *intention to be* is slightly different from Bratman's [2]. In Bratman's theory an *intention to be* must be joined with an *intention to do*, that is, if an agent intends that p then he necessarily intends to do something in order to achieve p. Thus, according to Bratman, when an agent intends that p,

[3]Note that we weaken C&L's negative condition $Bel_i\neg\phi$ to $\neg Bel_i\phi$, the reason being that $AGoal^{CL}_iF\phi$ is inconsistent, which is contrary to intuitions.

[4]The idea that an agent builds plans in order to satisfy his goals when he believes that the achievement of what he wants depends on him, is related with a particular conception of the way instrumental intentions are generated. We adhere here to Von Wright's conception of practical inference [10] according to whom practical reasoning is best captured by reasoning from an end to the necessary means to that end.

he already has a plan to achieve p. In our logic, the relation between the notion of *intention to be* and the notion of *plan* is weaker than Bratman's. Our *intention to be* only needs a *dependence belief* which is the immediate precursor of an *intention to do*. Indeed, when an agent wants p to be true and believes that his intervention is needed to produce p, he is *at the beginning* of a planning process which will yield an *intention to do*.

4. APPLICATION TO DELEGATION

Our aim in this section is to start from the motivation-based theory of delegation of Falcone & Castelfranchi (F&C) [4] and to show that their notion of delegation is intimately related with the notion of *intention to be* as defined in the previous section.[5]

F&C individuate two different kinds of delegation. First, they define *passive delegation* as the delegation based on exploitation, on the passive achievement by i of the task. Agent i just exploits in his plan a fully autonomous action of agent j. In fact, i has only to recognize the possibility that j will realize a certain action and that this action ensures that the goal of i will be satisfied. In this case agent i 'passively' awaits the satisfaction of his goals given his expectation that j will ensure it. Second, F&C define *active delegation* as the delegation based on the active indirect achievement by i of the task. Agent i not only exploits agent j's action for the achievement of his goals since he sees agent j's intervention as fundamental for this, but also acts in order to induce j to pursue the right course of action.

We provide the following formal definition of **passive delegation** in our STIT logic with mental states:

$$PassiveDel(i,j,\phi) \stackrel{\text{def}}{=}$$

$$\neg Bel_i \neg Stit_{AGT\setminus\{i\}} FStit_j\phi \wedge Pref_i FStit_j\phi \wedge \neg Bel_i\phi.$$

Thus, an agent i passively delegates the achievement of his goal that ϕ to agent j if and only if: (1) agent i wants to achieve ϕ by exploiting agent j (clause $Pref_i FStit_j\phi$); (2) agent i does not believe that ϕ is already achieved (clause $\neg Bel_i\phi$); (3) according to i's beliefs it is possible that there will be a moment where j will see to it that ϕ independently from what i does now (clause $\neg Bel_i \neg Stit_{AGT\setminus\{i\}} FStit_j\phi$). We can prove the following property of *passive delegation*.

THEOREM. $PassiveDel(i,j,\phi) \rightarrow \neg Int_i\phi \wedge \neg Int_i Stit_j\phi$.

This highlights the core nature of passive delegation: if agent i passively delegates the achievement of his goal that ϕ to agent j then agent i does not intend that ϕ and agent i does not intend that j sees to it that ϕ. Thus, in passive delegation agent i simply awaits for the realization of ϕ by (passively) exploiting agent j's action and without inducing j to cause ϕ.

Let us introduce the following formal definition of **active delegation**:

$$ActiveDel(i,j,\phi) \stackrel{\text{def}}{=}$$

$$Bel_i \neg Stit_{AGT\setminus\{i\}} FStit_j\phi \wedge Pref_i FStit_j\phi \wedge \neg Bel_i\phi \wedge$$

$$\neg Bel_i FStit_{AGT\setminus\{j\}}\phi.$$

Hence, an agent i actively delegates the achievement of his goal that ϕ to agent j if and only if: (1) agent i wants to achieve ϕ in the future by exploiting agent j (clause $Pref_i FStit_j\phi$); (2) agent i does not believe that ϕ is already achieved (clause $\neg Bel_i\phi$); (3) agent i believes that agent j will not see to it that ϕ independently from i's intervention ($Bel_i \neg Stit_{AGT\setminus\{i\}} FStit_j\phi$); (4) agent i does not believe that the future achievement of ϕ will be independent of what j will do (clause $\neg Bel_i FStit_{AGT\setminus\{j\}}\phi$).[6]

The following theorem captures the core nature of *active delegation*.

THEOREM. $ActiveDel(i,j,\phi) \rightarrow Int_i Stit_j\phi$.

Thus *active delegation* entails a specific kind of *intention to be*. If agent i actively delegates the achievement of his goal that ϕ to j then i has the intention that j sees to it that ϕ. Indeed in active delegation when i delegates to j the achievement of ϕ, he believes that he must induce and influence j to perform such action that will bring about ϕ. In *active delegation* the delegant either has decided to pursue a plan in order to induce the delegated agent to accomplish the task (i.e. the delegant intends to do something in order to induce the delegated agent to accomplish the task) or at least he is convinced that he must do something in order to induce the delegated agent to accomplish the task.

5. REFERENCES

[1] N. Belnap, M. Perloff, and M. Xu. *Facing the future: agents and choices in our indeterminist world*. Oxford, 2001.

[2] M. Bratman. *Intentions, plans, and practical reason*. Harvard University Press, Cambridge, 1987.

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

[4] R. Falcone and C. Castelfranchi. Towards a theory of delegation for agent-based systems. *Robotics and Autonomous Systems*, 24:141–157, 1998.

[5] R. Goldblatt. *Logics of Time and Computation, 2nd edition*. CSI Lecture Notes, Stanford, California, 1992.

[6] T. J. Norman and C. A. Reed. Group delegation and responsibility. In *Proc. of AAMAS'02*, 2002.

[7] M. D. Sadek. Dialogue acts are rational plans. In M. Taylor, F. Nel, and D. Bouwhuis, editors, *The structure of mutimodal dialogue*, pages 167–188, Philadelphia/Amsterdam, 2000. John Benjamins Publishing Company.

[8] J. Searle. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press, New York, 1983.

[9] W. Van der Hoek and M. Wooldridge. On the dynamics of delegation, cooperation, and control: A logical account. In *Proc. of AAMAS'05*, 2005.

[10] G. H. Von Wright. On so-called practical inference. *The Philosophical Review*, 15:39–53, 1972.

[5] Recently other researchers have developed formal systems for reasoning about delegation [6, 9]. Unfortunately, most of these approaches are poor in expressive power since they try to express the concept of delegation without referring to mental states of agents (goals, beliefs, intentions, etc...). Indeed, in our view delegation is intrinsically an intentional notion.

[6] This latter condition is given in order to capture the choice aspect involved in every act of delegation. Indeed agent i's belief that ϕ will not be achieved unless j will intervene and will do something, is a sufficient reason for i to decide to exploit j for achieving ϕ and to delegate to j the task of achieving ϕ.