Towards a Logical Theory of Coordination and Joint Ability

Hojjat Ghaderi
Department of Computer Science
University of Toronto
Toronto, ON M5S3G4, Canada
hojjat@cs.toronto.edu

Hector Levesque
Department of Computer Science
University of Toronto
Toronto, ON M5S3G4, Canada
hector@cs.toronto.edu

Yves Lespérance
Department of Computer Science and Engineering
York University
Toronto, ON M3J 1P3, Canada
lesperan@cs.yorku.ca

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Coherence and coordination; I.2.11 [**Distributed Artificial Intelligence**]: Multiagent systems

General Terms

Theory

1. INTRODUCTION

The coordination of teams of cooperating but autonomous agents is a core problem in multiagent systems research. A team of agents is *jointly able* to achieve a goal if despite any incomplete knowledge or even false beliefs that they may have about the world or each other, they still know enough to be able to get to a goal state, should they choose to do so. Unlike in the single-agent case, the mere existence of a working plan is not sufficient since there may be several incompatible working plans and the agents may not be able to choose a share that coordinates with the others'.

There is a large body of work in game theory [7] dealing with coordination and strategic reasoning for agents. The classical game theory framework has been very successful in dealing with many problems in this area. It can deal with cooperative teams of agents if we assign the same utility function to all team members. However, a major limitation of the framework is that it assumes there is a *complete specification* of the structure of the game including the agents' beliefs. It is also often assumed that this structure is common knowledge among agents. These assumptions often do not hold for team members, let alone for a third party attempting to reason about what the team members can do. In recent years, there has been a lot of work aimed at developing symbolic logics of games [8, 12] so that more incomplete and qualitative specifications can be dealt with. This can also lead to faster algorithms as sets of states that satisfy a property can be abstracted over in reasoning. However, this work has often incorporated very strong assumptions. Many logics of games like Coalition logic [8] and ATEL [12] ignore the issue of coordination within a coalition and assume that the coalition can achieve a goal if there exists a joint working plan/strategy profile that achieves the goal. This is only sufficient if we assume that the agents can communicate arbitrarily to agree on a joint plan. As well, most logics of games are propositional, which limits expressiveness.

In this paper, we develop a new first-order (with some higher-order features) logic framework to model the coordination of coalitions of agents based on the situation calculus [9]. Our formalization of joint ability avoids both of the pitfalls mentioned above: it supports reasoning on the basis of very incomplete specifications about the belief states of the team members and it ensures that team members do not have incompatible strategies. The formalization involves iterated elimination of dominated strategies [7, 1]. Each agent compares her strategies based on her private beliefs. Initially, they consider all available strategies possible. Then they eliminate strategies that are not as good as others given their beliefs about what strategies the other agents have kept. This elimination process is repeated until it converges to a set of preferred strategies for each agent. Joint ability holds if all combinations of preferred strategies succeed in achieving the goal.

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

2. THE FORMAL FRAMEWORK

The basis of our framework for joint ability is the situation calculus [6, 5]. The situation calculus is a predicate calculus language for representing dynamically changing domains. A *situation* represents a possible state of the domain. There is a set of initial situations corresponding to the ways the domain might be initially. The actual initial state of the domain is represented by the distinguished initial situation constant, S_0. The term $do(a, s)$ denotes the unique situation that results from an agent doing action a in situation s. Initial situations are defined as those that do not have a predecessor: $Init(s) \doteq \neg\exists a \exists s'. s = do(a, s')$. In general, the situations can be structured into a set of trees, where the root of each tree is an initial situation and the arcs are actions. The formula $s \sqsubseteq s'$ is used to state that there is a path from situation s to situation s'. Our account of joint ability will require some second-order features of the situation calculus, including quantifying over certain functions from situations to actions, that we call *strategies*. Predicates and functions whose values may change from situation to situation (and whose last argument is a situation) are called *fluents*. The effects of actions on fluents are defined using successor state axioms [9].

To represent belief of agents, we adopt Shapiro et al. [11]' logic of knowledge for multiple agents which is based on pos-

978-81-904262-7-5 (RPS) ©2007 IFAAMAS

sible world semantics in the situation calculus. The fluent $B(x, s', s)$ will be used to denote that in situation s, agent x thinks that situation s' might be the actual situation. Note that the order of the situation arguments is reversed from the convention in modal logic for accessibility relations. Belief is then defined as an abbreviation:[1]

$$Bel(x, \phi[now], s) \doteq \forall s'. B(x, s', s) \supset \phi[s'].$$

We require B to be serial and transitive, so that belief satisfies the modal system *weak S5*. We will also use

$$BW(x, \phi, s) \doteq Bel(x, \phi, s) \vee Bel(x, \neg\phi, s)$$

as an abbreviation for the agent x believing whether ϕ holds. When we need to represent knowledge we will require B to be reflexive as well (corresponding to the S5 modal system).

We now provide our proposed definition of joint ability. We assume there are N agents named 1 to N trying to achieve a common goal *Goal*. For space reasons, we only provide an English gloss of the formal definitions (see [2]).

- $JCan(s)$: Agents can jointly achieve the goal iff all combinations of their *preferred* strategies *work* together.
- $Works(\vec{\sigma}, s)$: Strategy profile $\vec{\sigma}$ works in situation s if there is a future situation where the common goal *Goal* holds and the strategies prescribe the actions to get there according to whose turn it is.
- $Pref(i, \sigma_i, s) \doteq \forall n.\, Keep(i, n, \sigma_i, s)$
  Agent i prefers strategy σ_i if it is *kept* for all levels n.
- *Keep* is defined inductively:
  - $Keep(i, 0, \sigma_i, s) \doteq Strategy(i, \sigma_i)$.
    At level 0, all strategies are kept.
  - $Keep(i, n+1, \sigma_i, s)$: For each agent i, the strategies kept at level $n+1$ are those kept at level n for which there is not a better one (σ_i' is better than σ_i if it is as good as σ_i while σ_i is not as good as it).
- $Strategy(i, \sigma_i)$: Strategies for an agent are functions from situations to actions such that the action is known to the agent whenever it is the agent's turn to act.
- $AsGoodAs(i, n, \sigma_i, \sigma_i', s)$: Strategy σ_i is as good as σ_i' for agent i if i believes that whenever σ_i' works with strategies kept by the rest of the agents so does σ_i.

As we will see in the examples next, the mere *existence* of a working strategy profile is not enough; the definition requires coordination among the agents in that *all* preferred strategies must work together.

3. A SIMPLE GAME SETUP

To illustrate our formalization of joint ability, we will employ the following simple setup involving only two agents, P and Q, one distinguished fluent F, and one distinguished action A. For simplicity, we assume that the agents act synchronously and in turn: P acts first and then they alternate. We assume that there is at least one other action A', and possibly more. All these actions are public (observed by both agents) and can always be executed. We assume for the purpose of this paper that there are no *preestablished* conventions that would allow agents to rule out or prefer strategies to others or to use actions as signals for coordination (e.g. similar to those used in the game of bridge).

The successor state axioms for the fluents are as follows:

- $F(do(a, s)) \equiv F(s)$.
  The fluent F is unaffected by any action.
- $turn(do(a, s)) = x \equiv$
  $x = P \wedge turn(s) = Q \;\vee\; x = Q \wedge turn(s) = P$.
  Whose turn it is to act alternates between P and Q.
- $B(x, s', do(a, s)) \equiv \exists s''. B(x, s'', s) \wedge s' = do(a, s'')$.
  This is a simplified version of the successor state axiom proposed by Scherl and Levesque. It is appropriate when actions have no preconditions, when there are no sensing (or knowledge-producing actions), and all actions are public to all agents.

We also include the following initial state axiom:
$Init(s) \supset turn(s) = P$. So, agent P gets to act first.
Let Σ_e denote the action theory containing the above axioms together with seriality and transitivity axioms for B. Also, let AX_{ReflxB} be the reflexivity axiom for B.

The sorts of goals we will consider will only depend on whether or not the fluent F held initially, whether or not P did action A first, and whether or not Q then did action A. Since there are $2 \times 2 \times 2$ options, and since a goal can be satisfied by any subset of these options, there are $2^8 = 256$ possible goals to consider. This does not mean, however, that there are only 256 possible games. We assume the agents can have beliefs about F and about each other. Since they may have beliefs about the other's beliefs about their beliefs and so on, there are, in fact, an infinite number of games. At one extreme, we may choose not to stipulate anything about the beliefs of the agents; at the other extreme, we may specify completely what each agent believes. In between, we may specify some beliefs or disbeliefs and leave the rest of their mental state open. For each such specification, and for each of the 256 goals, we may ask whether the agents are jointly able to achieve that goal, and whether they (*mutually*) *believe* this.

Example 1: Suppose we only stipulate that Q knows that P knows whether or not F holds. Consider a goal that is satisfied by P doing A and Q not doing A if F is true and P not doing A and Q doing A if F is false:

$$Goal(s) \doteq \exists s', a.\, Init(s') \wedge a \neq A \wedge [F(s') \wedge s = do(a, do(A, s')) \vee \neg F(s') \wedge s = do(A, do(a, s'))].$$

In this case, the two agents can achieve the goal: P will do the right thing since he knows whether F is true; Q will then do the opposite of P since he knows that P knows what to do. The action of P in this case behaves like a signal to Q.

Theorem 1. $\Sigma_e \cup AX_{ReflxB} \models$
$\forall s.\, [Init(s) \wedge Bel(Q, BW(P, F, now), s)] \supset JCan(s)$.

Note that we have left unspecified what Q believes about F. In fact, no matter what beliefs/disbeliefs he has about F they can still achieve the goal. Also, it follows immediately that common knowledge of the fact that P knows whether F holds implies common knowledge of joint ability. Interestingly, if we drop the reflexivity axiom and merely require

[1] Free variables are assumed to be universally quantified from outside. If ϕ has a single free situation variable, $\phi[t]$ denotes ϕ with that variable replaced by situation term t.

Q to *believe* that P knows whether or not F holds, then even if this belief is true, it is not sufficient to imply joint ability.

Example 2: Suppose again we stipulate that Q knows that P knows whether or not F holds. Consider a goal that is satisfied by P doing anything and Q not doing A if F is true and P doing anything and Q doing A if F is false:

$$\begin{aligned} Goal(s) \doteq \exists s', a, b.\ Init(s') \wedge \\ [F(s') \wedge b \neq A \wedge s = do(b, do(a, s')) \vee \\ \neg F(s') \wedge b = A \wedge s = do(b, do(a, s'))]. \end{aligned}$$

In a sense this is a goal that is easier to achieve than the one in Example 1 (i.e. $Goal_{Ex1}(s) \supset Goal_{Ex2}(s)$), since it does not require any specific action from P. Nonetheless, in this case, it would not follow that they can achieve the goal:

THEOREM 2. $\Sigma_e \cup AX_{ReflxB} \not\models \forall s. [Init(s) \wedge Bel(Q, BW(P, F, now), s)] \supset JCan(s)$.

Note that there are two profiles that the agents believe achieve the goal: (1) P does A when F holds and a non-A action otherwise, and Q does the opposite of P's action; (2) P does a non-A action when F holds and A otherwise, and Q does the same action as P. However, Q does not know which strategy P will follow and hence might choose an incompatible strategy. Therefore, although there are working joint plans, the existence of at least two incompatible preferred joint plans results in the lack of joint ability.

We can prove that if Q knows whether F holds they can achieve the goal as Q will know exactly what to do:

THEOREM 3. $\Sigma_e \cup AX_{ReflxB} \models Init(s) \wedge BW(Q, F, s) \supset JCan(s)$.

4. DISCUSSION

The definition of joint ability presented here is w.r.t. N agents all trying to achieve a common goal. It can be straightforwardly generalized to allow some agents to be outside of the coalition [2]. As mentioned in Section 1, there has been much recent work on developing symbolic logics of cooperation. In [13] the authors propose a model of cooperative problem solving and define joint ability by simply adapting the definition of single-agent ability, i.e. they take the existence of a joint plan that the agents mutually believe achieves the goal as sufficient for joint ability. They address coordination in the plan formation phase where agents negotiate to agree on a promising plan before starting to act. Coalition logic, introduced in [8], formalizes reasoning about the power of coalitions in strategic settings. It has modal operators corresponding to a coalition being able to enforce various outcomes. The framework is propositional and also ignores the issue of coordination *inside* the coalition. In a similar vein, van der Hoek and Wooldridge propose ATEL, a variant of alternating-time temporal logic enriched with epistemic relations in [12]. ATEL-based frameworks also ignore the issue of coordination inside a coalition. In [4], the authors acknowledge this shortcoming and address it by enriching the framework with extra cooperation operators. However, these operators nonetheless require either communication among coalition members, or a third-party choosing a plan for the coalition.

The issue of coordination using domination-based solution concepts has been throughly explored in game theory [7]. Our framework differs from these approaches, however, in a number of ways. Foremost, our framework not only handles incomplete information [3], but also incomplete *specifications* where some aspects of the world or agents including belief/disbelief are left unspecified. Since our proofs are based on entailment, they remain valid should we add more detail to the theory. Second, rather than considering utility functions, our focus is on goal achievability for a team. Moreover, we consider strict uncertainty and assume that no probabilistic information is available. Our framework supports a weaker form of belief (as in the weak S5 logic) and allows for false belief. Although our definition of joint ability resembles the notion of admissibility and iterated weak dominance in game theory [7, 1], the traditional definition of weak dominance cannot be used and an alternative approach for comparing strategies (as described in Section 2 is) is needed, one that is based on the private beliefs of each agent about the world and other agents and their beliefs.

In future work, we will work to generalize our framework in various ways. Supporting sensing actions and simple communication actions should be straightforwardly handled by revising the successor state axiom for belief accessibility as in [10, 11]. We will also examine how different ways of comparing strategies (the $AsGoodAs$ order) lead to different notions of joint ability, and try to identify the best. We will also evaluate our framework on more complex game settings. Finally, we will look at how our framework could be used in automated verification and multiagent planning.

5. REFERENCES

[1] A. Brandenburger and H. J. Keisler. Epistemic conditions for iterated admissibility. In *TARK '01*, pages 31–37, San Francisco, 2001. Morgan Kaufmann.

[2] H. Ghaderi. *A Logical Theory of Coordination and Joint Ability, In preparation*. PhD thesis, Department of Computer Science, University of Toronto, 2007.

[3] J. C. Harsanyi. Games with incomplete information played by Bayesian players. *Management Science*, 14:59–182, 1967.

[4] W. Jamroga and W. van der Hoek. Agents that know how to play. *Fundamenta Informaticae*, 63(2-3):185–219, 2004.

[5] H. Levesque, F. Pirri, and R. Reiter. Foundations for the situation calculus. *Electronic Transactions on Artificial Intelligence*, 2(3-4):159–178, 1998.

[6] J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, editors, *Machine Intelligence 4*, pages 463–502. 1969.

[7] M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, 1999.

[8] M. Pauly. A modal logic for coalitional power in games. *J. of Logic and Computation*, 12(1):149–166, 2002.

[9] R. Reiter. *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. The MIT Press, 2001.

[10] R. Scherl and H. Levesque. Knowledge, action, and the frame problem. *Artificial Intelligence*, 144:1–39, 2003.

[11] S. Shapiro, Y. Lesprance, and H. J. Levesque. Specifying communicative multi-agent systems. In W. Wobcke, M. Pagnucco, and C. Zhang, editors, *Agents and Multiagent systems- Formalisms, Methodologies, and Applications*, pages 1–14. Springer-Verlag, 1998.

[12] W. van der Hoek and M. Wooldridge. Cooperation, knowledge, and time: Alternating-time temporal epistemic logic and its applications. *Studia Logica*, 75(1):125–157, 2003.

[13] M. Wooldridge and N. R. Jennings. The cooperative problem-solving process. *Journal of Logic and Computation*, 9(4):563–592, 1999.