# A Logical Theory of Coordination and Joint Ability

**Hojjat Ghaderi** and **Hector Levesque**
Department of Computer Science
University of Toronto
Toronto, ON M5S 3G4, Canada
{hojjat,hector}@cs.toronto.edu

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

**Abstract**

A team of agents is jointly able to achieve a goal if despite any incomplete knowledge they may have about the world or each other, they still know enough to be able to get to a goal state. Unlike in the single-agent case, the mere existence of a working plan is not enough as there may be several incompatible working plans and the agents may not be able to choose a share that coordinates with those of the others. Some formalizations of joint ability ignore this issue of coordination within a coalition. Others, including those based on game theory, deal with coordination, but require a complete specification of what the agents believe. Such a complete specification is often not available. Here we present a new formalization of joint ability based on logical entailment in the situation calculus that avoids both of these pitfalls.

## 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 those of the others.

There is a large body of work in game theory (Osborne & Rubinstein 1999) 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. However, a major limitation of the framework is that it assumes that there is a *complete specification* of the structure of the game including the beliefs of the agents. It is also often assumed that this structure is common knowledge among the agents. These assumptions often do not hold for the 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 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 of its own. Many logics of games like Coalition Logic (Pauly 2002) and ATEL (van der Hoek & Wooldridge 2003) ignore the issue of coordination within a coalition and assume that the coalition can achieve a goal if there exists a 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 / strategy profile. In addition, 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. 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 (Osborne & Rubinstein 1999). Each agent compares her strategies based on her private beliefs. Initially, they consider all 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.

In the next section, we describe a simple game setup that we use to generate example games, and test our account of joint ability. Then, we present our formalization of joint ability in the situation calculus. We show some examples of the kind of ability results we can obtain in this logic. This includes examples where we prove that joint ability holds given very weak assumptions about the agents. Then, we discuss related work and summarize our contributions.

## A simple game setup

To illustrate our formalization of joint ability, we will employ a simple game setup that incorporates several simplifying assumptions. Many of them (*e.g.* only two agents, public actions, no communicative acts, goals that can be achieved with just two actions) are either not required by our formalization or are easy to circumvent; others are harder to get around. We will return to these in the Discussion section.

In our examples, there are two agents, $P$ and $Q$, one distinguished fluent $F$, and one distinguished action $A$. The agents act synchronously and in turn: $P$ acts first and then they alternate. There is at least one other action $A'$, and possibly more. All actions are public (observed by both agents) and can always be executed. 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 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 $A$.[1] 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 internal state open. For each such specification, and for each of the 256 goals, we may ask if the agents can jointly achieve the goal.[2]

**Example 1:** Suppose nothing is specified about the beliefs of $P$ and $Q$. Consider a goal that is satisfied by $P$ doing $A$ and $Q$ not doing $A$ regardless of $F$. In this case, $P$ and $Q$ can jointly achieve the goal, since they do not need to know anything about $F$ or each other to do so. Had we stipulated that $P$ believed that $F$ was true and $Q$ believed that $F$ was false, we would still say that they could achieve the goal despite the false belief that one of them has.

**Example 2:** Suppose we 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. 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 $Q$. Interestingly, if we merely require $Q$ to *believe* that $P$ knows whether or not $F$ holds, then even if this belief is true, it would not be sufficient to imply joint ability (specifically, in the case where it is true for the wrong reason; we will return to this).

**Example 3:** 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. In a sense this is a goal that is easier to achieve than the one in Example 2, since it does not require any specific action from $P$. Yet, in this case, it would not follow that they can achieve the goal. Had we additionally stipulated that $Q$ did not know whether $F$ held, we could be more definite and say that they definitely cannot jointly achieve this goal as there is nothing $P$ can do to help $Q$ figure out what to do.

**Example 4:** Suppose again we stipulate that $Q$ knows that $P$ knows whether or not $F$ holds. Consider a goal that is like in Example 3 but easier, in that it also holds if both agents do not do $A$ when $F$ is false. In this case, they can achieve the goal. The reason, however, is quite subtle and depends on looking at the various cases according to what $P$ and $Q$ believe. Similar to Example 2, requiring $Q$ to have true belief about $P$ knowing whether $F$ holds is not sufficient.

To the best of our knowledge, there is no existing formal account where examples like these and their variants can be formulated. We will return to this in the Related Work section. In the next section, we present a formalization of joint ability that handles the game setup above and much more based on entailment in the situation calculus.

## The formal framework

The basis of our framework for joint ability is the situation calculus (McCarthy & Hayes 1969; Levesque, Pirri, & Reiter 1998). 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: $\mathit{Init}(s) \doteq \neg\exists a \exists s'.\, s = do(a, s')$. 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 (Reiter 2001), which provide a succinct representation for both effect and frame axioms (McCarthy & Hayes 1969). To axiomatize a dynamic domain in the situation calculus, we use action theories (Reiter 2001) consisting of (1) successor state axioms; (2) initial state axioms, which describe the initial states of the domain including the initial beliefs of the agents; (3) precondition axioms, which specify the conditions under which each action can be executed (we assume here that all actions are always possible); (4) unique names axioms for the actions, and (5) domain-independent foundational axioms (we adopt the ones given in (Levesque, Pirri, & Reiter 1998) which accommodate multiple initial situations).

For our examples, we only need three fluents: the fluent $F$ mentioned in the previous section in terms of which goals are formulated, a fluent *turn* which says whose turn it is to act, and a fluent $B$ to deal with the beliefs of the agents.

[1] We use the term "goal" not in the sense of an agent's attitude, but as a label for the state of affairs that we ask whether the agents have enough information to achieve, *should they choose to do so*.

[2] We may also ask whether the agents *believe* or *mutually believe* that they have joint ability, but we defer this to later.

Moore (Moore 1985) defined a possible-worlds semantics for a logic of knowledge in the situation calculus by treating situations as possible worlds. Scherl and Levesque (Scherl & Levesque 2003) adapted this to Reiter's action theories and gave a successor state axiom for $B$ that states how actions, including sensing actions, affect knowledge. Shapiro et al. (Shapiro, Lespérance, & Levesque 1998) adapted this to handle the beliefs of multiple agents, and we adopt their account here. $B(x, s', s)$ 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:[3]

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

We will also use

$$\mathit{TBel}(x, \phi[\mathit{now}], s) \doteq \mathit{Bel}(x, \phi[\mathit{now}], s) \wedge \phi[s]$$

as an abbreviation for true belief (which we distinguish from knowledge formalized as a *KT45* operator, for reasons alluded to above in Example 2). Whenever we need knowledge and not merely true belief, we can simply add the following initial reflexivity axiom (called **IBR**) to the theory:

$$\mathit{Init}(s) \supset B(x, s, s).$$

Mutual belief among the agents, denoted by *MBel*, can be defined either as a fix-point or by introducing a new accessibility relation using a second-order definition. Common knowledge is then *MBel* under the **IBR** assumption.

Our examples use the following successor state axioms:

- $F(do(a, s)) \equiv F(s)$.
  The fluent $F$ is unaffected by any action.
- $\mathit{turn}(do(a, s)) = x \equiv$
  $x = P \wedge \mathit{turn}(s) = Q \vee x = Q \wedge \mathit{turn}(s) = P$.
  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. See the Discussion section for how it can be generalized.

The examples also include the following initial state axioms:

- $\mathit{Init}(s) \supset \mathit{turn}(s) = P$. So, agent $P$ gets to act first.
- $\mathit{Init}(s) \wedge B(x, s', s) \supset \mathit{Init}(s')$.
  Each agent initially knows that it is in an initial situation.
- $\mathit{Init}(s) \supset \exists s' B(x, s', s)$.
  Each agent initially has consistent beliefs.
- $\mathit{Init}(s) \wedge B(x, s', s) \supset \forall s''.\, B(x, s'', s') \equiv B(x, s'', s)$.
  Each agent initially has introspection of her beliefs.

The last two properties of belief can be shown to hold for all situations using the successor state axiom for $B$ so that belief satisfies the modal system *KD45* (Chellas 1980). If we include the **IBR** axiom, belief will satisfy the modal system *KT45*. Since the axioms above are universally quantified, they are known to all agents, and in fact are common knowledge. We will let $\Sigma_e$ denote the action theory containing the successor and initial state axioms above. All the examples in the next section will use $\Sigma_e$ with additional conditions.

[3]Free variables are assumed to be universally quantified from outside. If $\phi$ is a formula with a single free situation variable, $\phi[t]$ denotes $\phi$ with that variable replaced by situation term $t$. Instead of $\phi[\mathit{now}]$ we occasionally omit the situation argument completely.

## Our definition of joint ability

We assume there are $N$ agents named 1 to $N$. We use the following abbreviations for representing strategy[4] profiles:

- A vector of size $N$ is used to denote a complete strategy profile, e.g. $\vec{\sigma}$ for $\sigma_1, \sigma_2, \cdots, \sigma_N$.
- $\vec{\sigma}_{-_i}$ represents an incomplete profile with strategies for everyone except player $i$, i.e. $\sigma_1, \cdots, \sigma_{i-1}, \sigma_{i+1} \cdots, \sigma_N$.
- $\oplus_i$ is used to insert a strategy for player $i$ into an incomplete profile: $\vec{\sigma}_{-_i} \oplus_i \delta : \sigma_1, \cdots, \sigma_{i-1}, \delta, \sigma_{i+1} \cdots, \sigma_N$.
- $|_i$ is used to substitute the $i$th player's strategy in a complete profile: $\vec{\sigma}|_i \delta : \sigma_1, \cdots, \sigma_{i-1}, \delta, \sigma_{i+1} \cdots, \sigma_N$.

All of the definitions below are abbreviations for formulas in the language of the situation calculus presented above. The joint ability of $N$ agents to achieve $\phi$ is defined as follows:[5]

- $\mathit{JCan}(\phi, s) \doteq$
  $\forall \vec{\sigma}.\, [\bigwedge_{i=1}^{N} \mathit{Pref}(i, \sigma_i, \phi, s)] \supset \mathit{Works}(\vec{\sigma}, \phi, s)$.
  Agents $1 \cdots N$ can jointly achieve $\phi$ iff all combinations of their preferred strategies work together.
- $\mathit{Works}(\vec{\sigma}, \phi, s) \doteq \exists s''.\, s \sqsubseteq s'' \wedge \phi[s''] \wedge$
  $\forall s'.\, s \sqsubseteq s' \sqsubset s'' \supset$
  $\bigwedge_{i=1}^{N} (\mathit{turn}(s') = i \supset do(\sigma_i(s'), s') \sqsubseteq s'')$.
  Strategy profile $\vec{\sigma}$ works if there is a future situation where $\phi$ holds and the strategies in the profile prescribe the actions to get there according to whose turn it is.
- $\mathit{Pref}(i, \sigma_i, \phi, s) \doteq \forall n.\, \mathit{Keep}(i, n, \sigma_i, \phi, s)$
  Agent $i$ prefers strategy $\sigma_i$ if it is kept for all levels $n$.[6]
- *Keep* is defined inductively:[7]
  - $\mathit{Keep}(i, 0, \sigma_i, \phi, s) \doteq \mathit{Strategy}(i, \sigma_i)$.
    At level 0, all strategies are kept.
  - $\mathit{Keep}(i, n+1, \sigma_i, \phi, s) \doteq \mathit{Keep}(i, n, \sigma_i, \phi, s) \wedge$
    $\neg \exists \sigma_i'.\, \mathit{Keep}(i, n, \sigma_i', \phi, s) \wedge$
    $\mathit{GTE}(i, n, \sigma_i', \sigma_i, \phi, s) \wedge \neg \mathit{GTE}(i, n, \sigma_i, \sigma_i', \phi, s)$.
    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 ($\sigma_i'$ is better than $\sigma_i$ if it is as good as, i.e. greater than or equal to, $\sigma_i$ while $\sigma_i$ is not as good as it).
- $\mathit{Strategy}(i, \sigma_i) \doteq$
  $\forall s.\, \mathit{turn}(s) = i \supset \exists a.\, \mathit{TBel}(i, \sigma_i(\mathit{now}) = a, s)$.

[4]Strictly speaking, the $\sigma_i$'s are second-order variables ranging over functions from situations to actions. We use $\mathit{Strategy}(i, \sigma_i)$ to restrict them to valid strategies.

[5]In the Discussion section, we consider the case where there may be agents outside of the coalition of $N$ agents.

[6]The quantification is over the sort natural number.

[7]Strictly speaking, the definition we propose here is ill-formed. We want to use it with the second argument universally quantified (as in *Pref*). *Keep* and *GTE* actually need to be defined using second-order logic, from which the definitions here emerge as consequences. We omit the details for space reasons.

Strategies for an agent are functions from situations to actions such that the required action is known to the agent whenever it is the agent's turn to act.

- $GTE(i, n, \sigma_i, \sigma_i', \phi, s) \doteq Bel(i, \forall \vec{\sigma}_{-i}. ([\bigwedge_{j \neq i} Keep(j, n, \sigma_j, \phi, now) \wedge Works(\vec{\sigma}_{-i} \oplus_i \sigma_i', \phi, now)] \supset Works(\vec{\sigma}_{-i} \oplus_i \sigma_i, \phi, now)), s)$.
  Strategy $\sigma_i$ is as good as (Greater Than or Equal to) $\sigma_i'$ for agent $i$ if $i$ believes that whenever $\sigma_i'$ works with strategies kept by the rest of the agents so does $\sigma_i$.

These formulas define joint ability in a way that resembles the iterative elimination of weakly dominated strategies of game theory (Osborne & Rubinstein 1999) (see the Related Work section). As we will see in the examples to follow, 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.

## Formalizing the examples

As we mentioned, for each of the 256 possible goals we can consider various assumptions about the beliefs of the agents. In this section, we provide theorems for the goals of the four examples mentioned earlier. Due to lack of space we omit the proofs. Since there are only two agents, the previous definitions can be simplified. For better exposition, we use $g$ (possibly superscripted) to refer to strategies of the first agent (called $P$) and $h$ for those of the second (called $Q$).

### Example 1

For this example, the goal is defined as follows:
$\phi_1(s) \doteq \exists s'. Init(s') \wedge \exists a.\, a \neq A \wedge s = do(a, do(A, s'))$
Because the goal in this example (and other examples) is only satisfied after exactly two actions, we can prove that $Works(g, h, \phi_1, s)$ depends only on the first action prescribed by $g$ and the response prescribed by $h$:

**Lemma 1** $\Sigma_e \models Init(s) \supset Works(g, h, \phi_1, s) \equiv g(s) = A \wedge h(do(A, s)) \neq A.$

We can also show that $P$ will only prefer to do $A$, and $Q$ will only prefer to do a non-$A$ action. So, we have the following:

**Theorem 1** $\Sigma_e \models Init(s) \supset JCan(\phi_1, s).$

It then trivially follows that the agents can achieve $\phi_1$ despite having false beliefs about $F$:

**Corollary 1** $\Sigma_e \models [Init(s) \wedge Bel(P, \neg F, s) \wedge Bel(Q, F, s)] \supset JCan(\phi_1, s).$

We can also trivially show that the agents have mutual belief that joint ability holds:

**Corollary 2** $\Sigma_e \models Init(s) \supset MBel(JCan(\phi_1, now), s).$

### Example 2

For this example, the goal is defined as follows:
$\phi_2(s) \doteq \exists s', a.\, Init(s') \wedge a \neq A \wedge [F(s') \wedge s = do(a, do(A, s')) \vee \neg F(s') \wedge s = do(A, do(a, s'))].$

**Lemma 2** $\Sigma_e \models Init(s) \supset Works(g, h, \phi_2, s) \equiv [F(s) \wedge g(s) = A \wedge h(do(A, s)) \neq A \vee \neg F(s) \wedge g(s) \neq A \wedge h(do(g(s), s)) = A].$

We will also use the following definitions:

- $BW(x, \phi, s) \doteq Bel(x, \phi, s) \vee Bel(x, \neg\phi, s)$
  the agent believes whether $\phi$ holds.
- $TBW(x, \phi, s) \doteq TBel(x, \phi, s) \vee TBel(x, \neg\phi, s)$.

As mentioned in Example 2, $Q$'s having true belief about $P$ truly believing whether $F$ is not sufficient for joint ability:

**Theorem 2** $\Sigma_e \cup \{TBel(Q, TBW(P, F, now), S_0)\} \not\models JCan(\phi_2, S_0).$

This is because $TBel(Q, TBW(P, F, now), s)$ does not preclude $Q$ having a false belief about $P$, namely believing that $P$ believes that $F$ is false when in fact $P$ believes correctly that $F$ is true. To resolve this, we can simply add the **IBR** axiom. Another approach is to remain in the *KD45* logic but assert that $Q$'s belief about $P$'s belief of $F$ is correct:

$$BTBel(Q, P, F, s) \doteq [Bel(Q, TBel(P, F, now), s) \supset TBel(P, F, s)] \wedge [Bel(Q, TBel(P, \neg F, now), s) \supset TBel(P, \neg F, s)]$$

To keep our framework as general as possible, we take the second approach and add $BTBel(Q, P, F, s)$ whenever needed. With this, we have the following theorem:

**Theorem 3** $\Sigma_e \models [Init(s) \wedge BTBel(Q, P, F, s) \wedge TBel(Q, TBW(P, F, now), s)] \supset JCan(\phi_2, s).$

It follows from the theorem that $Q$'s *knowing* that $P$ *knows* whether $F$ holds is sufficient to get joint ability. More interestingly, it follows immediately that common knowledge of the fact that $P$ knows whether $F$ holds implies common knowledge of joint ability.

### Example 3

The goal for this example is easier to satisfy than the one in Example 2 (i.e. $\Sigma_e \models \phi_2(s) \supset \phi_3(s)$):
$\phi_3(s) \doteq \exists s', a, b.\, Init(s') \wedge [F(s') \wedge b \neq A \wedge s = do(b, do(a, s')) \vee \neg F(s') \wedge b = A \wedge s = do(b, do(a, s'))].$
Nonetheless, unlike in Example 2, it does not follow that the agents can achieve the goal (cf. theorem 3):

**Theorem 4** $\Sigma_e \not\models [BTBel(Q, P, F, S_0) \wedge TBel(Q, TBW(P, F, now), S_0)] \supset JCan(\phi_3, S_0).$

In fact, we can prove a stronger result that if $Q$ does not believe whether $F$ holds they cannot jointly achieve $\phi_3$:

**Theorem 5** $\Sigma_e \models \neg BW(Q, F, S_0) \supset \neg JCan(\phi_3, S_0).$

Note that there are two strategy profiles that the agents believe achieve $\phi_3$: (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 profiles, the existence of at least two incompatible kept profiles results in the lack of joint ability. We did not have this problem in Example 2 since profile (2) did not work there.

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

**Theorem 6** $\Sigma_e \models Init(s) \wedge TBW(Q, F, s) \supset JCan(\phi_3, s).$

**Example 4**

The goal here is easier than the ones in Examples 2 and 3:
$\phi_4(s) \doteq \exists s', a, b.\, \mathit{Init}(s') \wedge$
$\{F(s') \wedge b \neq A \wedge s = do(b, do(a, s')) \vee \neg F(s') \wedge$
$[s = do(A, do(A, s')) \vee b \neq A \wedge s = do(a, do(b, s'))]\}$.
Similarly to Example 2, we can show that if $Q$ has true belief about $P$ truly believing whether $F$ holds, then assuming $\mathit{BTBel}(Q, P, F, s)$, the agents can achieve the goal:

**Theorem 7** $\Sigma_e \models \mathit{Init}(s) \wedge \mathit{BTBel}(Q, P, F, s) \wedge \mathit{TBel}(Q, \mathit{TBW}(P, F, \mathit{now}), s) \supset \mathit{JCan}(\phi_4, s)$.

Note that there are many profiles that achieve $\phi_4$ (including profiles (1) and (2) mentioned in Example 3). Nonetheless, unlike in Example 3, we can prove by looking at various cases that the agents can coordinate (even if $\neg \mathit{BW}(Q, F, s)$ holds) by eliminating their dominated strategies.

From the above theorem, it follows that $Q$'s *knowing* that $P$ knows whether $F$ holds is sufficient to get joint ability. More interestingly, common knowledge of the fact that $P$ knows whether $F$ holds implies common knowledge of joint ability even though $Q$ may have incomplete beliefs about $F$.

## Properties of the definition

We now present several properties of our definition to show its plausibility in general terms. Let $\Sigma$ be an arbitrary action theory with a *KD45* logic for the beliefs of $N$ agents.

Our definition of ability is quite general and can be nested within beliefs. The consequential closure property of belief can be used to prove various subjective properties about joint ability. For example, to prove $\mathit{Bel}(i, \mathit{Bel}(j, \mathit{JCan}(\phi, \mathit{now}), \mathit{now}), S_0)$, it is sufficient to find a formula $\gamma$ such that $\Sigma \models \forall s.\, \gamma[s] \supset \mathit{JCan}(\mathit{now}, s)$ and $\Sigma \models \mathit{Bel}(i, \mathit{Bel}(j, \gamma, \mathit{now}), S_0)$.

One simple case where we can show that an agent believes that joint ability holds is when there is no need to coordinate. In particular, if agent $i$ has a strategy that she believes achieves the goal regardless of choices of other team members, then she believes that joint ability holds:[8]

**Theorem 8** $\Sigma \models [\exists \sigma_i \forall \vec{\sigma}_{-_i}.\, \mathit{Bel}(i, \mathit{Works}(\vec{\sigma}_{-_i} \oplus_i \sigma_i, \phi, \mathit{now}), S_0)] \supset \mathit{Bel}(i, \mathit{JCan}(\phi, \mathit{now}), S_0)$.[9]

(Example 3 with $\mathit{BW}(Q, F, s)$ is such a case: $Q$ believes that a strategy that says do $A$ when $F$ is false and do a non-$A$ action otherwise achieves the goal whatever $P$ chooses.) However, there are theories $\Sigma'$ such that even though agent $i$ has a strategy that always achieves the goal (regardless of choices of others) joint ability does not actually follow:

**Theorem 9** *There are $\Sigma'$ and $\phi$ such that*
$\Sigma' \cup \{\exists \sigma_i \forall \vec{\sigma}_{-_i} \mathit{Works}(\vec{\sigma}_{-_i} \oplus_i \sigma_i, \phi, S_0)\} \not\models \mathit{JCan}(\phi, S_0)$.

This is because agent $i$ might not know that $\sigma_i$ always works, and hence might keep other incompatible strategies as well.

Another simple case where joint ability holds is when there exists a strategy profile that every agent truly believes works, and moreover everyone believes it is impossible to achieve the goal if someone deviates from this profile:[10]

**Theorem 10** $\Sigma \models [\exists \vec{\sigma}.\, \mathit{ETBel}(\mathit{Works}(\vec{\sigma}, \phi, \mathit{now}), S_0) \wedge \forall \vec{\sigma}' \neq \vec{\sigma}.\, \mathit{EBel}(\neg \mathit{Works}(\vec{\sigma}', \phi, \mathit{now}), S_0)] \supset \mathit{JCan}(\phi, S_0)$.

It turns out that joint ability can be proved from even weaker conditions. $\mathit{ETBel}(\mathit{Works}(\vec{\sigma}, \phi, \mathit{now}), S_0)$ in the antecedent of Theorem 10 can be replaced by $\mathit{Works}(\vec{\sigma}, \phi, S_0) \wedge \bigwedge_{i=1}^{N} \neg \mathit{Bel}(i, \neg \mathit{Works}(\vec{\sigma}, \phi, \mathit{now}), S_0)$, i.e. $\mathit{Works}(\vec{\sigma}, \phi, S_0)$ holds and is consistent with the beliefs of the agents. This is because each agent $i$ will only prefer her share of $\vec{\sigma}$ (i.e. $\sigma_i$).

We can generalize the result in Theorem 10 if we assume there exists a strategy profile that is *known* by everyone to achieve the goal. Then, it is sufficient for every agent to *know* that their share in the profile is at least as good as any other available strategy to them, for *JCan* to hold:

**Theorem 11** $\Sigma \cup \{\mathit{IBR}\} \models [\exists \vec{\sigma}.\, \mathit{EBel}(\mathit{Works}(\vec{\sigma}, \phi, \mathit{now}), S_0) \wedge \forall \vec{\sigma}'.\, \bigwedge_{i=1}^{N} \mathit{Bel}(i, \mathit{Works}(\vec{\sigma}', \phi, \mathit{now}) \supset \mathit{Works}(\vec{\sigma}'|_i \sigma_i, \phi, \mathit{now}), S_0)] \supset \mathit{JCan}(\phi, S_0)$.

Another important property of joint ability is that it is non-monotonic w.r.t. the goal. Unlike in the single agent case, it might be the case that a team is able to achieve a strong goal while it is unable to achieve a weaker one (the goals in Examples 3 and 4 are an instance of this):

**Theorem 12** *There are $\Sigma'$, $\phi$, and $\phi'$ such that* $\Sigma' \models \phi(s) \supset \phi'(s)$ *but* $\Sigma' \not\models \mathit{JCan}(\phi, S_0) \supset \mathit{JCan}(\phi', S_0)$.

## Discussion

Although the game in this paper makes several simplifying assumptions for illustration purposes, many of them are not actually required by our account of joint ability or are easy to circumvent. For example, the formalization as presented here supports more than two agents and goals that are achieved after more than two actions. Although the game does not involve any sensing or communicative acts, our definition stays the same should we include such actions. We only need to revise the successor state axiom for belief accessibility as in (Scherl & Levesque 2003; Shapiro, Lespérance, & Levesque 1998). Similarly, actions that are not public can be handled with a variant axiom making the action known only to the agent who performed it. However, handling concurrent actions and actions with preconditions is more challenging.

Also, the definition of joint ability presented here assumes all $N$ agents are in the same coalition. It can be straightforwardly generalized to allow some agents to be outside of the coalition. Let $C$ be a coalition (i.e. a subset of agents $\{1, \cdots, N\}$). Since each agent $j \notin C$ might conceivably choose any of her strategies, agents inside the coalition $C$ must coordinate to make sure their choices achieve the goal regardless of the choices of the agents outside $C$. It turns out that a very slight modification to the definition of *Keep* is sufficient for this purpose. In particular, the definition of *Keep* for agents inside $C$ remains unchanged while for every agent $j \notin C$, we define $\mathit{Keep}(j, n, \sigma_j, s) \doteq \mathit{Strategy}(j, \sigma_j)$. Therefore, for every agent $j$ outside the coalition we have $\mathit{Pref}(j, \sigma_j, s) \equiv \mathit{Strategy}(j, \sigma_j)$.

[8] Note that, however, this does not imply that joint ability holds in the real world since $i$'s beliefs might be wrong.

[9] From here on, $\sigma$'s are intended to range over strategies.

[10] $\mathit{EBel}(\gamma, s) \doteq \bigwedge_{i=1}^{N} \mathit{Bel}(i, \gamma, s)$.

## Related work

As mentioned, there has been much recent work on developing symbolic logics of cooperation. In (Wooldridge & Jennings 1999) 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 (Pauly 2002), 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 (van der Hoek & Wooldridge 2003). Their framework also ignores the issue of coordination inside a coalition. In (Jamroga & van der Hoek 2004), the authors acknowledge this shortcoming and address it by enriching the framework with extra cooperation operators. 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 thoroughly explored in game theory (Osborne & Rubinstein 1999). Our framework differs from these approaches, however, in a number of ways. Foremost, our framework not only handles agents with incomplete information (Harsanyi 1967), but also it handles 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 assigning utility functions, our focus is on the achievability of a state of affairs by a team. Moreover, we consider strict uncertainty where probabilistic information may be unavailable. Our framework supports a weaker form of belief (as in *KD45* logic) and allows for false belief. Our definition of joint ability resembles the notion of admissibility and iterated weak dominance in game theory (Osborne & Rubinstein 1999; Brandenburger & Keisler 2001). Our work can be related to these by noting that every *model* of our theory with a *KT45* logic of belief can be considered as a partial extensive form game with incomplete information represented by a set of infinite trees each of which is rooted at an initial situation. We can add Nature as a player who decides which tree will be chosen as the real world and is indifferent among all her choices. Also, for all agents (other than Nature) we assign utility 1 to any situation that has a situation in its past history where $\phi$ is satisfied, and utility 0 to all other situations. However, since there are neither terminal nodes nor probabilistic information, the traditional definition of weak dominance cannot be used and an alternative approach for comparing strategies (as described in this paper) is needed, one that is based on the private beliefs of each agent about the world and other agents and their beliefs.

## Conclusion

In this paper, we proposed a logical framework based on the situation calculus for reasoning about the coordination of teams of agents. We developed a formalization of joint ability that supports reasoning on the basis of very incomplete specifications of the belief states of the team members, something that classical game theory does not allow. In contrast to other game logics, our formalization ensures that team members are properly coordinated. We showed how one can obtain proofs of the presence or lack of joint ability for various examples involving incomplete specifications of the beliefs of the agents. We also proved several general properties about our definitions.

In future work, we will consider some of the generalizations of the framework noted in the Discussion section. We will also examine how different ways of comparing strategies (the *GTE* 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.

## References


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

Chellas, B. 1980. *Modal logic: an introduction*. United Kingdom: Cambridge University Press.

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

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

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

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

Moore, R. 1985. A formal theory of knowledge and action. In Hobbs, J., and Moore, R., eds., *Formal Theories of the Commonsense World*, 319–358.

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

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

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

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

Shapiro, S.; Lespérance, Y.; and Levesque, H. 1998. Specifying communicative multi-agent systems. In Wobcke, W.; Pagnucco, M.; and Zhang, C., eds., *Agents and Multiagent systems*. Springer-Verlag. 1–14.

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

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