A Hybrid Approach to Teamwork

Paulo Trigo[1], **Helder Coelho**[2]

[1] Departamento de Engenharia de Electrónica e Telecomunicações e de Computadores
Instituto Superior de Engenharia de Lisboa, Portugal

[2]Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa, Portugal

ptrigo@deetc.isel.ipl.pt, hcoelho@di.fc.ul.pt

Abstract. *In the aftermath of a large-scale disaster, agents' decisions derive from self-interested (e.g. survival), common-good (e.g. victims' rescue) and teamwork (e.g. fire extinction) motivations. However, decision-theoretic models find it difficult to incorporate motivations, and mental-state models find it difficult to deal with uncertainty. We present an hybrid, CvI-JI, approach that combines: i) collective 'versus' individual (CvI) decisions, founded on the Markov decision process (MDP) quantitative evaluation of joint-actions, and ii) joint--intentions (JI) formulation of teamwork, founded on the belief-desire-intention (BDI) architecture of general mental-state based reasoning. Experiments show the CvI-JI performance's improvement during a policy learning process.*

1. Introduction

The agents that cooperate to mitigate the effects of a large-scale disaster, e.g. an earthquake or a terrorist incident, take decisions that follow two large behavioral classes: the individual (ground) activity and the collective (institutional) coordination of such activity. Additionally, agents are motivated to form teams and jointly commit to goals that supersede the individual capabilities. The collective 'versus' individual (CvI) decision model [Trigo et al. 2006], founded on the Markov decision process (MDP) framework, aims to conciliate the reciprocal influence of those two behavioral classes (collective and individual). Despite that, the CvI misses the agents' intentional stance toward team activity. On the other hand, the joint-intentions (JI) formulation of teamwork [Cohen and Levesque 1991], based on the belief-desire-intention (BDI) mental--state architecture, captures the agents' intentional stance, but misses the MDP domain--independent support for sequential decision-making in stochastic environments. Research on single-agent MDP-BDI hybrid approaches formulate the correspondence between the BDI plan and the MDP policy concepts [Simari and Parsons 2006] and empirically compares each model's performance [Schut et al. 2002]. Approaches to multi-agent MDP-BDI hybrid models often exploit BDI plans to improve MDP tractability, and use MDP to improve BDI plan performance [Tambe et al. 2005]. In this paper we take a different approach. Instead of exploring the plan-policy relationship, we focus on the relation between the BDI intention concept and the MDP *temporally abstract action* concept; we envisage an intention as an action that executes for time variable periods and, when terminated, yields a reward to the agent. We extend this view to the joint-intentions concept and integrate the resulting formulation in the 2-strata multilevel hierarchal CvI decision

model. The motivation for the hybrid CvI-JI model is to utilize the JI as an heuristic constraint that reduces the space of admissible joint-actions. The experimental results show the CvI-JI policy learning improvement in a partially observable environment.

The next section describes the CvI decision model and the section 3 outlines the JI concepts that are most relevant for our hybrid approach. Section 4 formulates the hybrid CvI-JI decision model, which is experimentally instantiated and evaluated in section 5. Finally, section 6 presents our conclusions and future research goals.

2. The collective 'versus' individual (CvI) decision model

The CvI decision model considers that the individual choice coexist with the collective choice and that coordinated behavior happens (is learned) from the prolonged relation (in time) of the choices exercised at both of those strata (individual and collective). Additionally, coordination is exercised on high level *cooperation tasks*, represented within an hierarchical task organization. The tasks' hierarchy is founded on the framework of *Options* [Sutton et al. 1999], which extends the MDP theory to include *temporally abstract actions* (variable time duration tasks, whose execution resorts to a subset of primitive actions).

2.1. The framework of Options

Formally, an MDP is a 4-tuple $\mathcal{M} \equiv \langle \mathcal{S}, \mathcal{A}, \Psi, P, R \rangle$ model of stochastic sequential decision problems, where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions, $\Psi \subseteq \mathcal{S} \times \mathcal{A}$ is the set of admissible state-action pairs, $R(s, a)$ is the expected reward when action a is executed at s, and $P(s' \mid s, a)$ is the probability of being at state s' after executing a at state s.

Given an MDP, an option $o \equiv \langle \mathcal{I}, \pi, \beta \rangle$, consists of a set of states, $\mathcal{I} \subseteq \mathcal{S}$, from which the option can be initiated, a policy, π, for the choice of actions and a termination condition, β, which, for each state, gives the probability that the option terminates when that state is reached. The computation of optimal value functions and optimal policies, $\pi^{\star}$, resorts to the relation between options and actions in a semi-Markov decision process (SMDP). The relation is that "any MDP with a fixed set of options is a SMDP" [Sutton et al. 1999]. Thus, all the SMDP learning methods can be applied to the case where temporally extended options are used in an MDP.

The option is an element of a multilevel hierarchy in which the policy of each option chooses among other lower level options. Thus, at each time step, the agent's decision is entirely among options, some of which persist for a single time step (primitive action or one-step option), and others are temporarily extended (multi-step option).

2.2. The CvI collective and individual strata

The individual stratum is simply a set of agents, Υ, each agent, $j \in \Upsilon$, having its particular capabilities described as an hierarchy of options. The CvI model admits agent heterogeneity (diverse option hierarchies), as long as all hierarchies have the same number of levels (depth), i.e., a similar temporal abstraction is used to design all hierarchies.

The collective stratum consists of a single agent (e.g. an institutional agent) that represents the whole set of individual stratum agents. The collective stratum agent cannot act on its own; its actions must be materialized through the individual stratum agents. The purpose of the collective stratum is to coordinate the individual stratum. Formally, at the

collective stratum, each action is specified as a *collective option*, $o_{\vec{o}} = \langle \mathcal{I}_{\vec{o}}, \pi_{\vec{o}}, \beta_{\vec{o}} \rangle$, where $\vec{o} = \langle o^1, \ldots, o^{|\Upsilon|} \rangle$ represents the simultaneous execution of option $o^j \equiv \langle \mathcal{I}^j, \pi^j, \beta^j \rangle$ by each agent $j \in \Upsilon$. The set of agents, Υ, defines an option space, $\vec{\mathcal{O}} \subseteq \mathcal{O}^1 \times \ldots \times \mathcal{O}^{|\Upsilon|}$, where $\mathcal{O}^j$ is the set of agent j options and each $o_{\vec{o}} \in \vec{\mathcal{O}}$ is a *collective option*. The $\vec{\mathcal{O}}$ decomposes into $\vec{\mathcal{O}}_d$ disjoint subsets, each containing only the *collective options* available at the, d, hierarchical level, where $0 < d \leq D$ and level-0 is the hierarchy root, at which there are no options to choose from, and level-D is the hierarchy depth. A level d policy, π_d, is implicitly defined by the SMDP $\mathcal{M}_d$ with state set $\mathcal{S}$ and action set $\vec{\mathcal{O}}_d$. The $\mathcal{M}_d$ solution is the optimal way to choose the level d individual policies which, in the long run, gather the highest collective reward.

The figure 1 illustrates the CvI decision model where the individual stratum (each *agent* ^{j} task hierarchy) has 3 levels and thus the collective stratum (represented by two, $\vec{o}_1$ and $\vec{o}_2$, *collective option* instances) contains 2 levels; at each level, the set of diamond ended arcs, links the *collective option* to each of its individual policies.

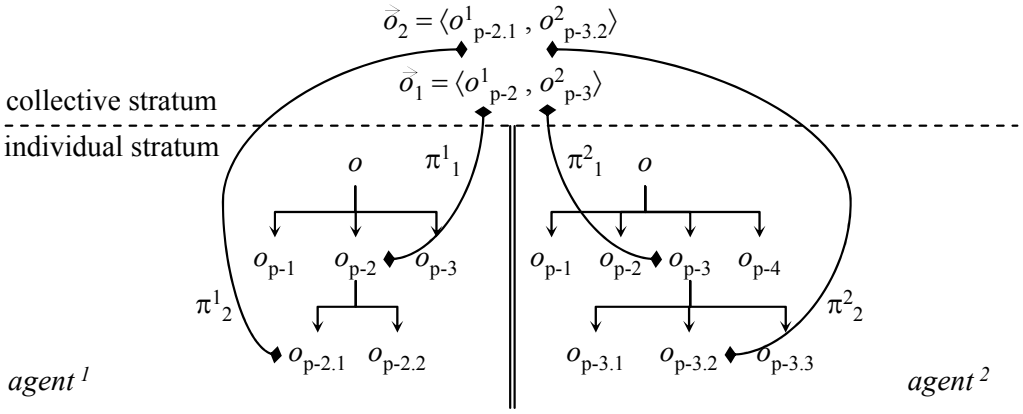

Figure 1. The CvI decision model and the links between strata (superscript j refers to *agent*^{j}; subscripts k and p-k refer to k hierarchical level and k tree path).

A centralized approach defines the $\mathcal{M}_d$ meta-policies and decides which individual policy to follow (i.e., decision-making is centralized in the collective stratum). Our CvI design follows a decentralized decision-making approach as it lets each agent choose whether to make a decision by itself or to ask the collective stratum for a decision. Such *decide-who-decides* (*d-w-d*) process is supported, at each hierarchical level d, by the value functions of the corresponding SMDP $\mathcal{M}_d$. The *d-w-d* represents the *importance* that an agent credits to collective and individual motivations, which is materialized as the ratio between, the maximum expected benefit in choosing a collective and an individual decision. A threshold, $\kappa \in [0, 1]$, enables to grade the focus from the collective to the individual stratum. Such regulatory mechanism enables the (human) designer to specify diverse social attitudes: ranging from self-interested to common-good motivated agents.

The formulation of the 2-stratum, CvI, multilevel hierarchical decision model and the definition of the inter-strata regulatory mechanism enabled to experimentally show how to explore the individual policy space in order to decrease the complexity of learning a coordination policy in a partially observable setting. We refer to [Trigo et al. 2006] for the comprehensive description of the CvI decision model.

2.3. The design of CvI agents

Given a set of agents, Υ, standing for the individual stratum, and an agent, υ, that impersonates the collective stratum, the design of a CvI instance is a 3-step process:

i. For each $j \in \Upsilon$, specify $\mathcal{O}^j$ — the set of options and its hierarchical organization.
ii. For each $j \in \Upsilon$, and from the agent υ perspective, identify the subset of *cooperation tasks*, $\mathcal{C}^j \subseteq \mathcal{O}^j$ — the most effective options to achieve coordination skills; the remaining options, $\mathcal{J}^j = \mathcal{O}^j - \mathcal{C}^j$, represent *purely individual tasks*.
iii. For each $j \in \Upsilon$, assign κ its regulatory value — where $\kappa = 0$ is a common-good motivated agent, $\kappa = 1$ is a self-interested attitude, and $\kappa \in]\,0, 1\,[$ embraces the whole spectrum between those two extreme decision motivations.

A simple, domain-independent design defines $\mathcal{C}^j$ (item ii above) as the set of multi--step options; hence $\mathcal{J}^j$ as the one-step options. Also, the highest hierarchical level(s) are usually effective to achieve coordination skills as they escape from getting lost in the confusion of lower level details. Our approach, at its current stage, requires a designer to specify domain-dependent collective and individual options (i.e., $\mathcal{C}^j$ and $\mathcal{J}^j$ sets).

3. The framework of joint-intentions (JI)

The precise semantics for the intention concept varies across the literature. An intention is often taken to represent an agent's internal commitment to perform an action, where a commitment is specified as a goal that persists over time, and a goal (often named as desire) is a proposition that the agent wants to get satisfied [Bratman 1990], [Cohen and Levesque 1990], [Rao and Georgeff 1995], [Wooldridge 2000]. An intention can also be taken to represent a linear plan that an agent has adopted to reach a state that the agent is committed to bring about [Georgeff and Ingrand 1989].

The framework of joint-intentions (JI) adopts the semantics of the "intention as a commitment to perform an action" and extends it to describe the concept of teamwork. A team is described as a set, of two or more agents, collectively committed to achieve a certain goal [Cohen and Levesque 1991]. The teamwork agents (those acting within a team) are expected to first form future-directed joint-intentions to act, keep those joint--intentions over time, and then jointly act. Formally, given a set of agents, Υ, a team is described as a 2-tuple $\mathcal{T} \equiv \langle\, \alpha,\ g \,\rangle$, where the team members are represented by $\alpha \subseteq \Upsilon$, and the team goal is g. In a team all members, α, are jointly committed to achieve the goal, g, while mutually believing that they are all acting towards that same goal. The teamwork terminates as soon as all members mutually believe that there exists at least one member that considers g as finished (achieved, impossible to achieve or irrelevant).

The CvI (cf. section 2) decision-theoretic model regards the JI approach as a way to reduce the *collective option* space exponentially in the number of team members. For example, given Υ agents, all with the same *cooperation tasks*, $\mathcal{C}$, there are at most $|\mathcal{C}|^{|\Upsilon|}$ admissible options to choose; during $\langle\, \alpha,\ g \,\rangle$ teamwork, that number reduces to $|\mathcal{C}|^{|\Upsilon|-|\alpha|}$ and such reduction motivates the formulation of the hybrid CvI-JI decision model.

4. The hybrid CvI-JI decision model

The formulation of the hybrid CvI-JI decision model addresses (in the next sections) two questions: i) how to specify, at design time, the JI using the CvI components?, and ii) how to integrate, at execution time, the JI specification in the CvI decision process?

4.1. Specification of JI using the CvI components

The JI describes teamwork in terms of goals which, in general, take multiple time periods until satisfaction. The CvI specifies decisions in terms of options which are *temporally abstract actions*. Therefore, a (team) goal corresponds to a (team) option. Given a goal, g, described as a proposition, φ, we formulate the corresponding option as $\langle \mathcal{I}, \pi, \beta \rangle$, where, $\mathcal{I}$ is the set of states such that $\neg\ \varphi$ is satisfied, $\beta(s) = 1$ if $s \in (\mathcal{S} - \mathcal{I})$ or $\beta(s) = 0$ otherwise, and π represents any policy to satisfy φ (i.e., to terminate the option).

The JI only requires agents to "keep the joint-intentions over time, and then jointly act"; it does not specify (the agent decides) when to terminate executing an ongoing task and effectively start acting to achieve a team goal. The CvI agents simultaneously execute their options, consequently, at a decision epoch, there may exist terminated and ongoing options. Thus, our hybrid CvI-JI option selection function distinguishes between two teamwork stages: i) the "ongoing task continue" while a team member executes another task, and ii) the "team option startup" when a team member starts executing the team option. Given a team member, j, and a team's option initiation set, $\mathcal{I}$, we define the ongoing states, $\mathcal{I}_{\text{ongo: } j} \subset \mathcal{I}$, where j is allowed to continue executing an ongoing task.

The JI assumes that once an agent commits to a team goal he will fulfil that commitment. The CvI is a stochastic model so we assume the possibility that an agent drops a previous commitment before actually starting to act as a team member. Given agent j we define the commitment probability, $p_{\text{commit: } j}$, that j meets his engagement.

The CvI-JI combines all the above elements (team option, ongoing set and commitment probability) into a single "teamwork design component" (*tdc*):

$$tdc \equiv \langle\, j,\ o,\ \mathcal{I}_{\text{ongo: } j},\ p_{\text{commit: } j} \,\rangle, \tag{1}$$

which describes, for each agent, $j \in \Upsilon$, and team option, o, the set of states, $\mathcal{I}_{\text{ongo: } j}$, where the agent may continue an ongoing task before committing to o, and the probability, $p_{\text{commit: } j}$, of effectively committing to o. The design of the *tdc* structure assumes that:

- a team option is always represented in more than one agent,
- an agent specifies a *tdc* instance for each team option he may get committed, and
- each $\mathcal{I}_{\text{ongo: } j}$ set is specified taking only the agent j local view of the environment.

The hybrid CvI-JI model describes, in the *tdc* instances, the domain-dependent teamwork knowledge which contributes to reduce the *collective option* space. Thus, CvI integrates JI as an heuristic filter (at collective stratum) that reifies the (human) designer domain knowledge. The next section integrates the heuristic filter in the decision process.

4.2. Integration of JI in the CvI decision process

The integration of the JI in the CvI decision process is designed, at the collective stratum, by modifying the CvI option selection process, which chooses, at each decision epoch, a level d *collective option*, $\vec{o}_d$ given a set of agents, $\mathcal{B}$, that request for a collective stratum decision. The algorithm 1 shows the option selection function, CHOOSEOPTION, and the inclusion of the two subroutines, APPLYFILTER-JI (cf. line 3) and UPDATEFILTER-JI (cf. line 5), that implement the CvI-JI integration.

The *getAdmissibleOptionSet* function (cf. evoked in algorithm 1, line 2) is exactly the same as in CvI; it evaluates the initiation set, $\mathcal{I}_{\vec{o}}$, of each *collective option*, $o_{\vec{o}}$, and

Algorithm 1 Choose option at the level d of the CvI collective stratum.

```
1  function CHOOSEOPTION( s, O_d, π_d, B )      ▷ B ≡ agents that request a decision
2      O_d' ← getAdmissibleOptionSet( s, O_d, B )   ▷ s ≡ collectively perceived state
3      O_d'' ← APPLYFILTER-JI( s, O_d', B )
4      o_d ← applyPolicy( s, O_d', O_d'', π_d )
5      UPDATEFILTER-JI( o_d, B )
6      return o_d
7  end function
```

returns the set, $\vec{\mathcal{O}}_d\,'$, of admissible options (given the perceived state, s, and the set of agents, $\mathcal{B}$, that requested a level d collective stratum decision). The *applyPolicy* function (cf. evoked in algorithm 1, line 4) chooses the next *collective option* to execute; the policy, π_d, is either predefined or follows some *explore-and-exploit* reinforcement learning method. We followed the learning approach and implemented a ϵ-greedy policy, which picks: i) a random admissible *collective option*, $o_{\vec{\sigma}} \in \vec{\mathcal{O}}_d\,'$, with probability ϵ, and ii) otherwise, picks the highest estimated action value *collective option*, at the current state, s, already considering the JI commitments (i.e., picks the $\max_{o_{\vec{\sigma}} \in \vec{\mathcal{O}}_d\,''} Q\,(\,s,\, o_{\vec{\sigma}}\,)$).

The algorithm 2, APPLYFILTER-JI function, shows the integration of JI commitments throughout the manipulation of the *tdc* instances (cf. expression 1). The first part of the function (cf. lines 2 to 10) determines the set *tdc* instances, TDC', that are consistent with the current situation. The second part (cf. lines 11 to 16) restricts the *collective options* to those that are compatible (all $o_{\vec{\sigma}}$ components match) with the team options of all $tdc \in TDC'$; the remaining *collective options* are discarded.

Algorithm 2 Apply JI filter to reduce the set of admissible *collective options*.

```
1  function APPLYFILTER-JI( s, O_d', B )
2      TDC' ← ∅
3      for each tdc ∈ TDC do                    ▷ TDC ≡ set of active tdc elements
4          if ( tdc.j ∈ B ) ∧ ( s_[j] ∉ tdc.I_ongo:j ) then   ▷ s_[j] ≡ agent j perceived state
5              if random ≤ tdc.p_commit:j then
6                  TDC' ← TDC' ∪ { tdc }
7              end if
8              TDC ← TDC − { tdc }
9          end if
10     end for
11     O_d'' ← ∅
12     for each o_σ ∈ O_d' do
13         if o_σ is compatible with TDC' then
14             O_d'' ← O_d'' ∪ { o_σ }
15         end if
16     end for
17     return O_d''                             ▷ O_d'' = O_d' when TDC' = ∅
18 end function
```

The algorithm 3, UPDATEFILTER-JI function, implements a strategy to define,

at each decision epoch, the set of active *tdc* elements (cf. expression 1 and *TDC* set in algorithm 2, line 3). We implemented a simple strategy where each agent "is available to commit to a team option as long as he is not already a team member"; the *TDC* set is updated according to that strategy, for all agents, at each decision epoch.

Algorithm 3 Update the set, *TDC*, of "teamwork design component" (*tdc*) elements.

```
1  function UPDATEFILTER-JI( o⃗_d, B )
2      teamOption ← false
3      for each tdc ∈ D_TDC do          ▷ D_TDC ≡ set of "design time" tdc elements
4          if ¬ teamOption then
5              o ← tdc.o                ▷ o ≡ a team option
6          end if
7          for each ag ∈ ϒ do
8              if ( o⃗_d[ag] ≠ o ) ∧ ( o⃗_d[tdc.j] = o ) ∧ ( ag ∈ B ) ∧ ( ag ≠ tdc.j ) then
9                  TDC ← TDC ∪ { tdc }   ▷ TDC ≡ globally initialized with ∅
10                 if ¬ teamOption then
11                     teamOption ← true
12                 end if
13             end if
14         end for
15     end for
16 end function
```

5. Experiment specification and results

We implemented the CvI-JI decision model and tested it in a multi-agent taxi environment: "a maze like grid inhabited by taxis (agents), passengers and sites". The original single-agent taxi problem was described as follows: "a passenger appears at a site and wish to be transported to another site; a taxi goes to the origin site of the passenger, pick up the passenger, go to its destination site and drop down the passenger" [Dietterich 2000]. A first multi-agent extension was described as follows: "there are multiple passengers and multiple taxis; the taxis may transport several passengers at the same time; a site may have several passengers, each with its own destination" [Trigo et al. 2006].

We further extend the previous multi-agent taxi problem, in order to enforce teamwork behavior, as follows: "there are some predefined sites where passengers only accept to be transported all together (as in a family); at those sites a taxi may not pick up more than one passenger (as if he was carrying a large luggage)". Those sites are named *teamwork sites* because taxis must work as a team to transport all passengers at the same time.

The environment is collectively observable as each taxi does not perceive the other taxis' locations, but their combined observations determine a sole world state. The goal of the individual stratum is to learn how to execute tasks (e.g. how to navigate to a site and when to pick up a passenger). The goal of the collective stratum is to learn to coordinate those individual tasks as to minimize the resources (time) to satisfy the passengers' needs.

We defined 3 different CvI-JI configurations, where a configuration is an assignment to all $j \in \Upsilon$ of the same $p_{\text{commit:}\,j} \in \{ 0, \frac{1}{2}, 1 \}$ value. Thus, we have:

- *never JI*, when $p_{\text{commit: } j} = 0$,
- *sometimes JI*, when $p_{\text{commit: } j} = \frac{1}{2}$, and
- *always JI*, when $p_{\text{commit: } j} = 1$.

The setup used for all experiments is: 5×5 grid, 4 sites $\mathcal{S}_b = \{ b_1, b_2, b_3, b_4 \}$, 2 taxis $\mathcal{S}_t = \{ t_1, t_2 \}$, 3 passengers $\mathcal{S}_{psg} = \{ psg_1, psg_2, psg_3 \}$, and a single *teamwork site* $b_{\text{tw}} \in \mathcal{S}_b$. The primitive actions, available to each taxi, are `pick`, `put`, `move(m)`, where `m` $\in \{$ `N`, `E`, `S`, `W` $\}$ are the cardinal directions, and the `wait` action (added to the original taxi problem) to support the agent's synchronization (e.g. at *teamwork sites*).

The learning of the policy at the collective stratum occurs simultaneously with the learning of each agent's policy at the individual stratum. The results of the experiments (cf. section 5.4) show the hybrid CvI-JI performance improvement of the collective stratum learning process, when compared with the pure CvI (i.e., *never JI*) approach.

5.1. JI specification

The JI specification consists in the set of predefined *tdc* instances. The *tdc* instance is defined, for each taxi (agent) $t_j \in \mathcal{S}_t$ as $\langle t_j, b_{\text{tw}}, \mathcal{I}_{\text{ongo: } t_j}, p_{\text{commit: } t_j} \rangle$. The $\mathcal{I}_{\text{ongo: } t_j}$ specifies the following ongoing state set: i) the taxi, t_j, already transports a passenger, or ii) there is a passenger to pick up at t_j current location. The $p_{\text{commit: } t_j}$ is assigned the value 0, $\frac{1}{2}$ or 1, respectively for the *never JI*, *sometimes JI* or *always JI* experiment configuration.

5.2. Individual stratum specification

The taxi observation, $\omega = \langle x, y, psg_1, psg_2, psg_3 \rangle$, represents its own (x, y)-position and passenger, $psg_i = \langle loc_i, dest_i, orig_i \rangle$, status where $loc_i \in \mathcal{S}_b \cup \mathcal{S}_t \cup \{ t_{1acc}, t_{2acc} \}$ (t_{1acc} means that taxi j accomplished delivery), $dest_i \in \mathcal{S}_b$, and $orig_i \in \mathcal{S}_b$.

The rewards provided to a taxi are: i) 20 for delivering a passenger, ii) -10 for illegal `pick` or `put`, iii) -12 for any illegal `move` action in a *teamwork site*, and iv) -1 for any other action, including moving into walls and picking more than one passenger in a *teamwork site*.

The task hierarchy is composed of a `root` option and a `navigate(b)` option for each `b` $\in \mathcal{S}_b$. Therefore, each agent holds an option hierarchy with 3 levels, where `root` is the level-zero option, `navigate(b)`, `pick`, `put` and `wait` are the level-one options and `move(m)` are the level-two one-step options (defined for each `navigate(b)`). We refer to [Trigo et al. 2006] for the full specification of the option hierarchy.

5.3. Collective stratum specification

The collective stratum holds the combined observations $s = \langle t_1, t_2, psg_1, psg_2, psg_3 \rangle$ of all agents, where t_j is the (x, y)-position of agent j. Our approach to the reward is to consider that agents equitably contribute to the current world state. Thus, the collective reward is defined as the sum of rewards provided to each agent; our purpose is to maximize the long run collective reward. The level-one *collective options* specification considers:

- $\mathcal{C} = \{$ `navigate(b)` for all $b \in \mathcal{S}_b \} \cup \{$ `wait` $\} \cup \{$ *indOp* $\}$, and
- $\mathcal{J} = \{$ `pick`, `put` $\}$,

where *indOp* is a special option that represents $\mathcal{J}$ at the collective stratum.

Within this experimental toy world, an individual agent perceives 52,428,800 states, and the collective stratum contains 1,310,720,000 states. Each individual decision considers 6 options, while for the collective stratum there are 36 *collective options*.

The decision-making of the collective stratum resorts, at each state, to the expected future value of each admissible *collective option*, whereas such evaluation is only acquired (learned) after the evidence (reward) gathered via the materialization (execution) of each decision. Hence, the experiments capture some of the complexity of the decision-making process that aims to achieve coordinated behavior in a disaster response environment.

5.4. Experimental results

The purpose of our experiments is to measure the influence of the JI integration in the CvI model. The performance of the learning process is used as the evaluation criterium and it is measured as the cumulative reward, gathered at the collective stratum, during an whole experiment. Each experiment executes for 700 episodes. An episode always starts with 2 passengers in the *teamwork site* and the third passenger in another site. Each episode terminates as soon as all passengers reach their destination. Policy learning uses a temporal difference approach (SMDP Q-learning [Bradtke and Duff 1995], [Sutton et al. 1999]) with the ϵ-greedy strategy previously described (cf. section 4.2). Each experiment starts with $\epsilon = 0.15$ and, after the first 100 episodes, ϵ decays 0.004 every each 50 episodes.

We ran 3 experiments, one for each CvI-JI configuration. The figure 2 shows that the *never JI* configuration exhibits the worst performance; it is about 6.5% worse than *always JI* and about 12% worse than *sometimes JI*; such difference remains almost uniform throughout the whole experiment. The *sometimes JI* reveals an unexpected behavior while, around episode 300, it starts to outperforms *always JI*.

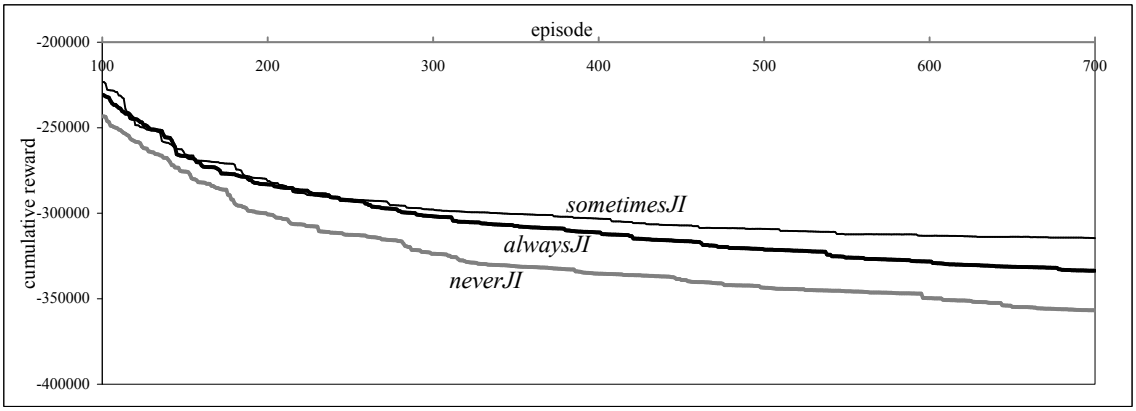

Figure 2. The influence of JI in the performance of the learning process.

An insight on these results is that the JI teamwork specific knowledge is exploited by the collective stratum, without compromising the exploration (search for novelty) that is required by the learning process. The unexpected result is that the capability of not fulfilling a previous teamwork commitment (cf. *sometimes JI*) enables to find improvements over the fully reliable commitment attitude (cf. *always JI*).

6. Conclusions and future work

In this paper, we have identified a series of relations between the 2-strata decision--theoretic CvI approach and the joint-intentions (JI) mental-state based reasoning. We have extended CvI by exploring the algorithmic aspects of the CvI-JI integration. Such integration represents our novel contribution to a multi-agent hybrid decision model within a reinforcement learning framework. The initial experimental results, of the CvI-JI model, sustain the hypothesis that the JI heuristic reduction of the action space improves the process of learning a policy to coordinate multiple agents. An interesting conclusion is that, taking into account our preliminary results, the stochastic commitment concept suggests investigating the hypothesis that not fulfilling a commitment (at a specific state) is an opportunity to find an alternative that, in the long run, is globally better than teamwork.

This work represents the ongoing steps in a line of research that aims to develop agents that participate in the decision-making process that occurs in the response to a large-scale disaster. Future work will apply the CvI-JI in a a simulated disaster response environment [Kitano and Tadokoro 2001] and will explore teamwork (re)formation strategies [Trigo and Coelho 2005] at the collective stratum.

Acknowledgments. Research was supported by PRODEP III 5.3/13/03 and LabMAG.

References

Bradtke, S. and Duff, M. (1995). Reinforcement learning methods for continuous-time Markov decision problems. In *Proceedings of Advances in Neural Information Processing Systems*, volume 7, pages 393–400. The MIT Press.

Bratman, M. (1990). What is intention? In *Intentions in Communication*, pages 15–31. MIT Press, Cambridge, MA.

Cohen, P. and Levesque, H. (1990). Intention is choice with commitment. *Artificial Intelligence*, 42(2–3):213–261.

Cohen, P. and Levesque, H. (1991). Teamwork. *Noûs, Special Issue on Cognitive Science and Artificial Intelligence*, 25(4):487–512.

Dietterich, T. (2000). Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, 13:227–303.

Georgeff, M. and Ingrand, F. (1989). Decision-making in an embedded reasoning system. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI-89)*, pages 972–978, Detroit, USA.

Kitano, H. and Tadokoro, S. (2001). RoboCup Rescue: A grand challenge for multi-agent systems. *Artificial Intelligence Magazine*, 22(1):39–52.

Rao, A. and Georgeff, M. (1995). BDI agents: From theory to practice. In *Proceedings of the First International Conference on Multiagent Systems*, pages 312–319, San Francisco, USA.

Schut, M., Wooldridge, M., and Parsons, S. (2002). On partially observable MDPs and BDI models. In *Foundations and Applications of Multi-Agent Systems*, volume 2403 of *Lecture Notes in Computer Science*, pages 243–260. Springer-Verlag.

Simari, G. and Parsons, S. (2006). On the relationship between MDPs and the BDI architecture. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-06)*, pages 1041–1048, Hakodate, Japan. ACM Press.

Sutton, R., Precup, D., and Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1–2):181–211.

Tambe, M., Bowring, E., Jung, H., Kaminka, G., Maheswaran, R., Marecki, J., Modi, P., Nair, R., Okamoto, S., Pearce, J., Paruchuri, P., Pynadath, D., Scerri, P., Schurr, N., and Varakantham, P. (2005). Conflicts in teamwork: Hybrids to the rescue. In *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-05)*, pages 3–10. ACM Press.

Trigo, P. and Coelho, H. (2005). The multi-team formation precursor of teamwork. In *Progress in Artificial Intelligence, EPIA-05*, volume 3808 of *Lecture Notes in Artificial Intelligence*, pages 560–571. Springer-Verlag.

Trigo, P., Jonsson, A., and Coelho, H. (2006). Coordination with collective and individual decisions. In *Advances in Artificial Intelligence, IBERAMIA/SBIA 2006*, volume 4140 of *Lecture Notes in Artificial Intelligence*, pages 37–47. Springer-Verlag.

Wooldridge, M. (2000). *Reasoning About Rational Agents*, chapter Implementing Rational Agents. The MIT Press.