ECAI 2008
M. Ghallab et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-891-5-413

A hybrid approach to multi-agent decision-making

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

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, current decision-theoretic models are either purely individual or purely collective and find it difficult to deal with motivational attitudes; on the other hand, mental-state based models find it difficult to deal with uncertainty. We propose a 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. The CvI-JI evaluation explores the performance's improvement during the process of learning a coordination policy in a partially observable stochastic domain.

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 their individual capabilities [8]. Despite such motivation, communication is usually insufficient to ensure that decision-making is supported by a single and coherent world perspective. The communication constraint causes the decision-making process to evolve simultaneously, both at the collective (common--good) and at the individual (self-interested) strata, sometimes in a conflicting manner. For instance, an ambulance searches for a policy to rescue a perceived civilian, while the ambulance command center, when faced with a global view of multiple injured civilians, searches for a policy to decide which ambulance should rescue which civilian.

However, despite the intuition on a 2-strata decision process, research on multi-agent coordination often proposes a single model that amalgamates both strata and searches for optimality within that model. The approaches based on the multi-agent Markov decision process (MMDP) [1] are *purely collective* and centralized, thus too complex to coordinate while requiring unconstrained communication. The multi-agent semi-Markov decision process (MSMDP) [7], although decentralized, requires each individual agent to represent the whole decision space (states and actions) which may become very large, thus causing the individual policy learning to be slow and highly dependent on up-to-date information about the decisions of all other agents. The game-theoretic approach requires an agent to compute the utility of all combinations of actions executed by all other agents (payoff matrix), which is then used to search for Nash equilibria (where no agent increases his payoff by unilaterally changing his policy); thus, if several equilibria exist, agents may adhere to purely individual policies never being pulled by a collective perspective.

The multi-agent collective 'versus' individual (CvI) decision model [15], which is founded on the semi-Markov decision process (SMDP) framework, is neither *purely collective* nor *purely individual* and explores the explicit separation of concerns between both (collective and individual) decision strata while aiming to conciliate their reciprocal influence. Despite that, the CvI misses the agents' intentional stance toward team activity. On the other hand, the joint--intentions (JI) formulation of teamwork [5], based on the belief--desire-intention (BDI) mental-state architecture [9, 16], 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 hybrids formulates the correspondence between the BDI plan and the MDP policy concepts [11] and empirically compares each model's performance [10]. Multi--agent MDP-BDI hybrid models often exploit BDI plans to improve MDP tractability, and use MDP to improve BDI plan selection [13].

In this paper, instead of exploring the MDP-BDI policy-plan relation, we focus on the link between the BDI intention concept and the MDP *temporally abstract action* concept [12]. We see 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 hierarchical CvI decision model. Thus, the CvI-JI is a hybrid approach that combines the MDP *temporally abstract action* concept and the BDI mental-state architecture. The motivation for the hybrid CvI-JI model is to use the JI as a heuristic constraint that reduces the space of admissible MDP joint-actions, thus enabling to escalate the problems' dimension. The experiments show the CvI-JI learning improvement in a partially observable environment.

2 THE CvI DECISION MODEL

The premise of the CvI decision model is that the individual choice coexists 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). Coordination is exercised on high level, hierarchically organized *cooperation tasks*, founded on the framework of *Options* [12], which extends the MDP theory to include *temporally abstract actions* (variable time duration tasks, whose execution resorts to 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, $\mathcal{A}$

[1] GuIAA/LabMAg; DEETC, ISEL - Instituto Superior de Engenharia de Lisboa, Lisbon, Portugal, email: ptrigo@deetc.isel.ipl.pt
[2] LabMAg; DI, FCUL - Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal, email: hcoelho@di.fc.ul.pt

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): "any MDP with a fixed set of options is a SMDP" [12]. Thus, all the SMDP learning methods can be applied to the case where temporally extended options are used in an MDP. The options define a multilevel hierarchy where the policy of an option chooses among other lower level options. At each time, the agent's decision is entirely among options; some persist for a single time step (primitive action or one--step option), 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, $j \in \Upsilon$, with its capabilities described as a hierarchy of options. The collective stratum is an agent (e.g. institutional) that cannot act on its own (its actions are executed by the individual stratum agents) and its purpose is to coordinate the individual stratum. Formally, at the collective stratum, each action is defined 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 with the *collective options* available at the, d, hierarchical level, where $0 < d \leq D - 1$ and level-0 is the root 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, gathers the highest collective reward.

The CvI structure. Figure 1 illustrates the CvI structure, where the individual stratum (each *agent*^{j}) is a 3-level hierarchy and thus the collective stratum (the two, $\vec{o}_1$ and $\vec{o}_2$, *collective option* instances) is a 2-level hierarchy; at each level, the set of diamond ended arcs, links the *collective option* to each of its individual policies.

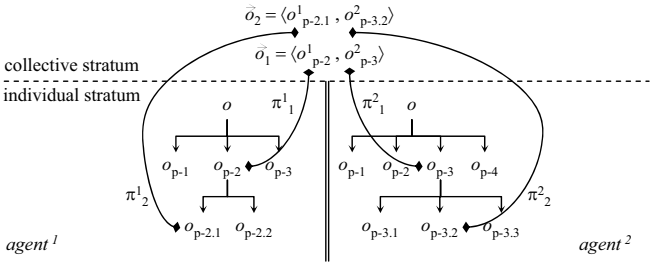

Figure 1. The CvI structure and inter-strata links (superscript j refers to *agent*^{j}; subscripts k and p-k refer to k hierarchical level and k tree path).

The CvI dynamics. At each decision epoch, *agent*^{j} gets the partial perception, ω^j, and *decide-who-decides* (*d-w-d*), i.e., the *agent*^{j} either: i) chooses an option $o^j \in \mathcal{O}^j$, or ii) requests, the collective stratum, which replies with an option, o^j, decision. The *d-w-d* process represents the *importance*, that an agent credits to each stratum, defined as the ratio between, the maximum expected benefit in choosing a collective and an individual decision. The expected benefit is given, at each hierarchical level-d, by the value functions of the corresponding SMDP $\mathcal{M}_d$. A threshold, $\kappa \in [\,0, 1\,]$, focus-grades between collective and individual strata, thus enabling the (human) designer to specify diverse social attitudes: ranging from common--good ($\kappa = 0$) to self-interested ($\kappa = 1$) motivated agents. The CvI is a decentralized model as each agent decides whether to make a decision by itself or to ask the collective layer for a decision. The comprehensive description of the CvI model refers to [15].

2.3 The design of CvI agents

Given the individual stratum set of agents, Υ, and a collective stratum agent, υ, 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 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 lower level details.

3 THE JOINT-INTENTIONS (JI) MODEL

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; an intention can also represent a plan that an agent has adopted to reach or a state that the agent is committed to bring about [3, 4, 9, 16].

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 [5]. 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).

4 THE HYBRID CvI-JI DECISION MODEL

Given the CvI (cf. section 2) decision-theoretic model we regard 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. The next sections address 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 Specify JI using the CvI components

The teamwork goal. 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 where $\neg\, \varphi$ is satisfied, $\beta(\, s\,) = 1$ if $s \in (\, \mathcal{S} - \mathcal{I}\,)$ or $\beta(\, s\,) = 0$ otherwise, and π is any policy to satisfy φ (i.e., to terminate the option).

The teamwork commitment. The JI only requires agents to "keep the joint-intentions commitment over time, and then jointly act". It is up to the agent to decide when to terminate executing an ongoing task and effectively start acting to achieve the team goal. Thus, being jointly committed to a goal, g, does not imply immediate action toward that same goal, g. For example, two ambulances may jointly commit to the same disaster while one of them is executing an action (e.g., delivering an injured civilian); as soon as the ongoing task is terminated, the ambulance starts acting towards the team goal. Therefore, our CvI-JI formulation assumes that, at each decision epoch, an agent may establish a JI while still acting to satisfy another intention (either individual or JI). Thus, at each instant, an agent may have an ongoing activity and also one (at most) established JI. Our approach enables teamwork decisions to be asynchronous; agents do not need to wait, for each others' option termination, before committing to a JI. Our hybrid CvI-JI option selection function distinguishes two teamwork stages: i) the "ongoing task continue" when an agent decides to establish a JI (becomes a team member) even though the agent still executes some other task, and ii) the "team option startup" when a team member decides to start executing the team option. Given a team member, j, a team option, o, and its 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 while jointly committed to achieve the team option, o.

The teamwork reconsideration. 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 teamwork design component. The CvI-JI combines all the above (team option, ongoing set and commitment probability) into a "teamwork design component" $tdc_j \equiv \langle\, o^j, \mathcal{I}_{\text{ongo: } j}, p_{\text{commit: } j} \,\rangle$, which describes, for agent, $j \in \Upsilon$, and team option, $o^j \in \mathcal{O}^j$, the set of states, $\mathcal{I}_{\text{ongo: } j}$, where the agent may continue an ongoing task before start executing o^j, and the probability, $p_{\text{commit: } j}$, of effectively committing to o^j. The design of the *tdc* structure assumes that: i) a team option is always represented in more than one agent, ii) a tdc_j is specified for each team option that j may get committed, and iii) the$\mathcal{I}_{\text{ongo: } j}$ specification considers the j's environment local view. The CvI-JI model describes, via *tdc*, 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 Integrate 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 perceived state, s, and 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.

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

```
1 function CHOOSEOPTION( s, O_d, π_d, B )
2     O_d' ← getAdmissibleOptionSet( s, O_d, B )
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
```

The *getAdmissibleOptionSet* function (cf. algorithm 1, line 2) is the same as in CvI; evaluates $\mathcal{I}_{\vec{o}}$ of each *collective option*, $o_{\vec{o}}$, and returns the set, $\vec{\mathcal{O}}_d\,'$, of admissible options (given the perceived s and the set of agents, $\mathcal{B}$, that requested a level d collective stratum decision). The *applyPolicy* function (cf. 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*, $o_{\vec{o}} \in \vec{\mathcal{O}}_d\,'$, 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 $\max_{o_{\vec{o}} \in \vec{\mathcal{O}}_d\,''} Q(\, s,\, o_{\vec{o}}\,)$).

The algorithm 2, APPLYFILTER-JI function, shows the integration of JI commitments throughout the manipulation of the *tdc* instances. The set of goals that call for teamwork effort are represented by the global *TDC* set (cf. line 3) which is initially empty. The first part (cf. lines 2 to 10, algorithm 2) determines the *TDC'* set of admissible *tdc* from agents that requested for a level d collective stratum decision. The teamwork reconsideration concept (cf. section 4.1) is represented by the possibility of discarding a previously established and currently admissible JI (cf. algorithm 2, line 5). The second part (cf. lines 11 to 16, algorithm 2) restricts the *collective options* to those that are compatible (all $o_{\vec{o}}$ components match) with the team options of all *tdc* $\in$ *TDC'*; the remaining *collective options* are discarded.

The algorithm 3, UPDATEFILTER-JI function, describes the strategy used, at each decision epoch, to select a team goal and to find the set of agents that are available to commit to that team goal (i.e., select a goal, g, and find the set, $\alpha \subseteq \Upsilon$, of agents available to form a team $\mathcal{T} \equiv \langle\, \alpha,\ g \,\rangle$). The implemented strategy simply selects the first admissible team goal and assumes that each agent "is available to commit to a team goal as long as he is not already a team member". The *TDC* set is updated (cf. algorithm 3) according to that strategy, for all agents, at each decision epoch.

5 EXPERIMENTS AND RESULTS

We propose the *teamwork taxi coordination problem* that extends the previous *taxi coordination problem* [6, 15] and enforces the team-

Algorithm 2 Apply JI to reduce *collective options*' admissible set.

```
1  function APPLYFILTER-JI( s, O_d', B )
2      TDC' ← ∅
3      for each tdc ∈ TDC do
4          if ( s_[j] ∉ tdc.I_ongo: j ) ∧ ( tdc.j ∈ B ) then
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
```

Algorithm 3 Strategy to update the set, *TDC*, containing the selected team goal and the agents available for a JI.

```
1  function UPDATEFILTER-JI( o_d, B )
2      teamOption ← false
3      for each tdc ∈ D_TDC do    ▷ D_TDC ≡ designed 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 ( o_d[ag] ≠ o ) ∧ ( o_d[tdc.j] = o ) ∧
9                 ( ag ∈ B ) ∧ ( ag ≠ tdc.j ) then
10                 TDC ← TDC ∪ { tdc }
11                 if ¬ teamOption then
12                     teamOption ← true
13                 end if
14             end if
15         end for
16     end for
17 end function
```

work behavior, as follows: "passengers appears at an origin site and wants to get transported to a destination site; there are some predefined sites where passengers only accept to get transported all together (as in a family)"; those sites are named *teamwork sites* as taxis must work as a team to transport all passengers at the same time.

The experimental setup is given by: i) a 5×5 grid, ii) 4 sites, $\mathcal{S}_b = \{ b_1, b_2, b_3, b_4 \}$, iii) 2 taxis, $\mathcal{S}_t = \{ t_1, t_2 \}$, iv) 3 passengers, $\mathcal{S}_{psg} = \{ psg_1, psg_2, psg_3 \}$, and v) a single, $b_{\text{tw}} \in \mathcal{S}_b$, *teamwork site*. 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 supports the agent's synchronization (at *teamwork sites*). The problem is partially observable as a taxi does not perceive the other taxis' locations; it is collectively observable as the combination of all individual observations determines a sole world state.

We defined 3 different CvI-JI configurations, each assigning all $j \in \Upsilon$ the same $p_{\text{commit: } j} \in \{ 0, \frac{1}{2}, 1 \}$ value. Therefore, we define: i) *never JI*, when $p_{\text{commit: } j} = 0$, ii) *sometimes JI*, when $p_{\text{commit: } j} = \frac{1}{2}$, and iii) *always JI*, when $p_{\text{commit: } j} = 1$.

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 the individual tasks to minimize the resources (time) to satisfy passengers' needs.

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 is specified as a set of predefined *tdc* instances. The *tdc* instance is defined, for each taxi (agent) $t_j \in \mathcal{S}_t$ as $\langle b_{\text{tw}}, \mathcal{I}_{\text{ongo: } t_j}, p_{\text{commit: } t_j} \rangle$. The b_{tw} is the *teamwork site*. 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

Each taxi's observation, $\omega = \langle x, y, psg_1, psg_2, psg_3 \rangle$, is its (x, y)-position and passenger, $psg_i = \langle loc_i, dest_i, orig_i \rangle$, status where $loc_i \in \mathcal{S}_b \cup \mathcal{S}_t \cup \{ t_{1_{acc}}, t_{2_{acc}} \}$ ($t_{1_{acc}}$ means that taxi j accomplished delivery), $dest_i \in \mathcal{S}_b$, and $orig_i \in \mathcal{S}_b$. Therefore, the state space, perceived by each taxi, is describe by a total of $5 \times 5 \times (8 \times 4 \times 4)^3 = 52{,}428{,}800$ states.

The taxi capability is a 3-level hierarchy, where `root` is the multi-step level-zero option, `navigate(b)` is the multi-step level--one option, `pick`, `put` and `wait` are the one-step level-one options and `move(m)` are the level-two one-step options (for each `navigate(b)`); a total of 5 multi-step options and 7 one-step actions. The taxi is not equipped with any explicit definition of its goal; also, it does not hold any internal representation of the maze grid. The taxi j decision is based solely on the information available at each decision epoch: i) its perception, ω^j, and ii) the immediate reward provided by the last executed one-step action.

The immediate taxi rewards 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*.

5.3 Collective stratum specification

The collective stratum perceives $s = \langle t_1, t_2, psg_1, psg_2, psg_3 \rangle$ which combines all the individual stratum partial observations, where t_j is the (x, y)-position of agent j. Therefore, the collective stratum state space is describe by $(5 \times 5)^2 \times (8 \times 4 \times 4)^3 = 1{,}310{,}720{,}000$ states. The collective stratum chooses mainly among multi-step options, so we specify: i) $\mathcal{C} = \{$ `navigate(b)` for all $b \in \mathcal{S}_b \} \cup \{$ `wait` $\} \cup \{$ *indOp* $\}$, and $\mathcal{J} = \{$ `pick`, `put` $\}$, where *indOp* is an implicit option representing $\mathcal{J}$ at the collective stratum. The *indOp* option gives place to a *ping-pong* decision scenario between strata, whenever an agent chooses to "request for a collective stratum decision" and the collective stratum replies: "decide yourself but consider only your *purely individual tasks*". Hence, the decision forwards back to the agent (via *indOp*) raising a second opportunity for the agent "to choose an option in $\mathcal{J}$". The *ping-pong* effect, while giving a second decision opportunity, does not increase the communication between strata and reduces, to $|\mathcal{J}|$, the individual decision space.

We assume that agents equitably contribute to the current state. Thus, the collective reward is the sum of rewards provided to each agent; our purpose is to maximize the long run collective reward.

5.4 The CvI-JI experimental evaluation

Our experiments evaluate the influence of the JI integration in the CvI model, by measuring the learning process performance (quantified as the collective stratum cumulative reward). An episode starts with 2 passengers in the *teamwork site* and the third passenger in another site; the episode terminates as soon as all passengers reach their destination; each experiment executes for 700 episodes. Policy learning follows the SMDP Q-learning [2, 12] approach with the ϵ-greedy strategy (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. Figure 2 shows that the *never JI* configuration exhibits the worst performance; about 6.5% worse than *always JI* and about 12% worse than *sometimes JI*; the 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*.

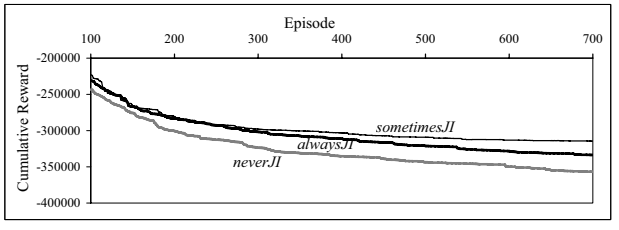

Figure 2. The influence of JI in the performance of the learning process.

An insight on these results is that the JI teamwork heuristic is exploited by the collective stratum, without compromising the exploration (search for novelty) that is required by the learning process. Somehow unexpected was that, being able not to fulfill a previous teamwork commitment (cf. *sometimes JI*), enables to find improvements over the fully reliable commitment attitude (cf. *always JI*).

The CvI-JI enables continuous (non interrupted) flow of decision--making and task execution activities. Such asynchronous process opens a time space between the instant the agent establishes a JI and the instant the agente actually begins acting to achieve the JI. The possibility to reconsider a commitment, just before actually start acting, explores alternatives to teamwork. The ability to drop a pre--established JI enables to find individual activity in state points where the heuristic approach (JI) would suggest a teamwork approach. Results (cf. figure 2) show that the exploration of individual policies combined with the heuristic teamwork approach enables to improve the process of learning a coordination policy.

The experiment's dimension. In this experiment, an agent perceives 52,428,800 states, and the collective stratum contains 1,310,720,000 states. Each decision considers 6 individual options and 36 *collective options*. Hence, this experimental world captures some of the complexity of the decision-making process that aims to achieve coordinated behavior in a disaster response environment.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we identified a series of relations between the 2-strata decision-theoretic CvI approach and the joint-intentions (JI) mental--state based reasoning. We 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 teamwork reconsideration concept suggests investigating the hypothesis that not fulfilling a commitment (at a specific state) is an opportunity to find an alternative path that, in the long run, is globally better than teamwork.

This work describes the ongoing research steps to construct 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 [8] and will explore teamwork (re)formation strategies [14] at the collective stratum.

ACKNOWLEDGEMENTS

This research was partially supported by LabMAg FCT R&D unit.

REFERENCES

[1] Craig Boutilier, 'Sequential optimality and coordination in multiagent systems', in *Proceedings of the Sixteenth International Joint Conferences on Artificial Intelligence (IJCAI-99)*, pp. 478–485, (1999).
[2] Steven Bradtke and Michael Duff, 'Reinforcement learning methods for continuous-time Markov decision problems', in *Proceedings of Advances in Neural Information Processing Systems*, volume 7, pp. 393–400. The MIT Press, (1995).
[3] Michael Bratman, 'What is intention?', in *Intentions in Communication*, 15–31, MIT Press, Cambridge, MA, (1990).
[4] Philip Cohen and Hector Levesque, 'Intention is choice with commitment', *Artificial Intelligence*, **42**(2–3), 213–261, (1990).
[5] Philip Cohen and Hector Levesque, 'Teamwork', *Noûs, Cognitive Science and Artificial Intelligence*, **25**(4), 487–512, (1991).
[6] Thomas Dietterich, 'Hierarchical reinforcement learning with the MAXQ value function decomposition', *Journal of Artificial Intelligence Research*, **13**, 227–303, (2000).
[7] Mohammad Ghavamzadeh, Sridhar Mahadevan, and Rajbala Makar, 'Hierarchical multi-agent reinforcement learning', *Autonomous Agents and Multi-Agent Systems*, **13**(2), 197–229, (2006).
[8] Hiroaki Kitano and Satoshi Tadokoro, 'RoboCup Rescue: A grand challenge for multi-agent systems', *AI Magazine*, **22**(1), 39–52, (2001).
[9] Anand Rao and Michael Georgeff, 'BDI agents: From theory to practice', in *Proceedings of the First International Conference on Multi-agent Systems*, pp. 312–319, San Francisco, USA, (1995).
[10] Martijn Schut, Michael Wooldridge, and Simon Parsons, 'On partially observable MDPs and BDI models', in *Foundations and Applications of Multi-Agent Systems*, volume 2403 of *LNCS*, 243–260, Springer-Verlag, (2002).
[11] Gerardo Simari and Simon Parsons, '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)*, pp. 1041–1048, Hakodate, Japan, (2006). ACM Press.
[12] Richard Sutton, Doina Precup, and Satinder Singh, 'Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning', *Artificial Intelligence*, **112**(1–2), 181–211, (1999).
[13] Milind Tambe, E. Bowring, H. Jung, Gal Kaminka, R. Maheswaran, J. Marecki, P. Modi, Ranjit Nair, S. Okamoto, J. Pearce, P. Paruchuri, David Pynadath, P. Scerri, N. Schurr, and Pradeep Varakantham, 'Conflicts in teamwork: Hybrids to the rescue', in *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-05)*, pp. 3–10. ACM Press, (2005).
[14] Paulo Trigo and Helder Coelho, 'The multi-team formation precursor of teamwork', in *Progress in Artificial Intelligence, EPIA-05*, volume 3808 of *LNAI*, 560–571, Springer-Verlag, (2005).
[15] Paulo Trigo, Anders Jonsson, and Helder Coelho, 'Coordination with collective and individual decisions', in *Advances in Artificial Intelligence, IBERAMIA/SBIA 2006*, volume 4140 of *LNAI*, 37–47, Springer-Verlag, (2006).
[16] Michael Wooldridge, *Reasoning About Rational Agents*, chapter Implementing Rational Agents, The MIT Press, 2000.