Coordination with Collective and Individual Decisions

Paulo Trigo[1], Anders Jonsson[2], Helder Coelho[3]

Dep. Eng. Elect. Telec. e Comp. at Instituto Superior de Engenharia de Lisboa, Portugal.[1]
Departamento de Tecnología at Universidad Pompeu Fabra, Barcelona, Spain.[2]
Departamento de Informática at Faculdade de Ciências, Universidade de Lisboa, Portugal.[3]
[1]ptrigo@deetc.isel.ipl.pt, [2]anders.jonsson@upf.edu, [3]hcoelho@di.fc.ul.pt

Abstract. The response to a large-scale disaster, e.g. an earthquake or a terrorist incident, urges for low-cost policies that coordinate sequential decisions of multiple agents. Decisions range from collective (common good) to individual (self-interested) perspectives, intuitively shaping a two-layer decision model. However, current decision theoretic models are either purely collective or purely individual and seek optimal policies. We present a two-layer, collective versus individual (CvI) decision model and explore the tradeoff between cost reduction and loss of optimality while learning coordination skills. Experiments, in a partially observable domain, test our approach for learning a collective policy and results show near-optimal policies that exhibit coordinated behavior.

1 Introduction

A coordination policy recommends decisions expected to bring agents to cooperate on a collective task (e.g. disaster mitigation) or at the very least, not to pursue conflicting strategies (e.g. compete to rescue a victim). The search for a coordination policy that responds to a large-scale disaster, such as an earthquake, is a process beyond individual skills where optimality is non-existent or too expensive to compute [9]. In many cases, communication is insufficient to ensure a single and coherent world perspective. Such communication constraints cause decision-making to occur both at collective (common good) and individual (self-interested) layers, sometimes in a conflicting manner. For instance, an ambulance searches for a policy to rescue a civilian, while the ambulance command center, when faced with a global view of multiple injured civilians, searches for a policy that avoids conflicts and decides which ambulance should rescue which civilian. However, despite the intuition on a two-layered decision, research on multi-agent coordination often proposes a single model that amalgamates those layers and searches for optimality within that model. A centralized model, e.g. the multi-agent Markov decision process (MMDP) [2], builds a purely collective world perspective that is too complex to coordinate and which requires unconstrained communication capability. In a decentralized model, e.g. the multi-agent semi-Markov decision process (MSMDP) [7], an agent makes decisions based on information about the decisions of other agents. If communication does not occur frequently, such information quickly becomes outdated. In addition, the state space of individual agents may become very large, causing learning to be slow. Also,

there exist game theoretic approaches that require each 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 [10] (where no agent increases his payoff by unilaterally changing his policy); thus, when several equilibrium exist, agents may adhere to individual policies that are not pulled by a collective perspective.

Therefore, our distinctive research hypothesis is: i) a two-layer model intuitively represents the decision-making that occur in complex domains, and ii) a model that includes collective and individual decisions, enables an agent to decide whether a decision should be made at the collective or at the individual level.

We formulate the *collective versus individual*, CvI, two-layer decision model. Each layer is a multilevel decision hierarchy supported on the framework of options [13], which extends the theory of reinforcement learning to include temporally abstract actions. The layers are linked via the formulation of a *collective layer option* over individual options and we use an *inform-request* scheme to communicate between layers; a *regulatory mechanism* provides the means to choose the layer of a decision.

The CvI decision model was experimentally tested in a collectively observable environment [11] (i.e. partially observable, where the combined partial views determines a sole state). The experimental results support our conjectural hypothesis and show how to reduce the learning cost yielding a near-optimal and admissible (from our qualitative analysis) coordination policy.

2 The Framework of Options

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

The framework of options [13] is founded on the MDP theory and extends the action concept to take variable amounts of time and frames the policy notion into such an extended action. Formally, given an MDP, an option $o \equiv \langle \mathsf{I}, \pi, \beta \rangle$, consists of a set of states, $\mathsf{I} \subseteq \mathsf{S}$, 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, π^*, 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" [13]. Thus, all the SMDP learning methods can be applied to the case where temporally extended options are used in an MDP.

The option representation provides one level of abstraction on top of primitive actions. However, a primitive action, a, also corresponds to an option that is available whenever a is available, that always terminates after one time step and that selects a everywhere. This uniformity defines a multilevel hierarchy in which the policy of an option chooses among other lower-level options. Thus, the agent's decision at each time step is entirely among options, some of which persist for a single time step (one-step option), and others are temporarily extended (multi-step option).

3 The CvI Collective and Individual Layers

The CvI collective layer formulation assumes that agents may be given different option hierarchies (heterogeneous agents), all hierarchies having the same number of levels (depth), i.e. a similar temporal abstraction is used to design all hierarchies.

We first recall (from [12]) the multi-option, $\bar{o}$, formulation: at each time instant the set of agents, Υ, concurrently executes a $|\Upsilon|$-tuple of options, $\bar{o} = \langle o^1, \ldots, o^{|\Upsilon|} \rangle$, such that each agent $j \in \Upsilon$ is executing option $o^j \equiv \langle I^j, \pi^j, \beta^j \rangle$, for j=1,…,$|\Upsilon|$.

In a partially observable environment, an agent may only observe part of the state. Let S^j be the set of partial states observed by agent j, and let H^j be the set of partial states hidden to j. The complete set of states is $S = S^j \times H^j$. For each option available to j, the initiation set is $I^j \subseteq S^j$, which means that the option is available whenever a state $s \in (I^j \times H^j) \subseteq S$, since for agent j the hidden part of the state is irrelevant.

We now formulate (over $\bar{o}$) our *collective layer option*, $o_{\bar{o}}$, concept: at the collective layer, an option has the form $o_{\bar{o}} \equiv \langle I_{\bar{o}}, \pi_{\bar{o}}, \beta_{\bar{o}} \rangle$ such that,

- $I_{\bar{o}} = \cap_{j=1..|\Upsilon|} (I^j \times H^j)$, i.e. $o_{\bar{o}}$ is admissible at any state $s \in S$ whenever each component of $\bar{o}$ is admissible at *s*,
- $\pi_{\bar{o}} = \langle \pi^1, \ldots, \pi^{|\Upsilon|} \rangle$, where π^j is the policy of agent $j \in \Upsilon$ while executing the component o^j of $\bar{o}$. Therefore, $\pi_{\bar{o}}$ (the policy of $o_{\bar{o}}$) is simply to follow the several, $|\Upsilon|$, individual policies,
- $\beta_{\bar{o}} = \tau(\beta^1, \ldots, \beta^{|\Upsilon|})$ where β^j is the termination condition of the component o^j of option $\bar{o}$. Since each o^j component may have a different duration, the τ function represents the termination scheme of the multi-option $\bar{o}$. In our model we always use the $\tau_{continue}$ scheme [12], where $\bar{o}$ terminates as soon as any o^j terminates but without dropping commitments (i.e. non terminating options keep executing); other referenced schemes are τ_{any}, τ_{all} [12] and τ_{change} [1].

The set of agents, Υ, defines an option space, $\bar{O} \subseteq O^1 \times \ldots \times O^{|\Upsilon|}$, where O^j is the set of options specified for agent j and each $o_{\bar{o}} \in \bar{O}$ is a collective layer option. Thus, the number of $o_{\bar{o}}$ options may grow exponentially with the number of agents, e.g. if all agents have the same, O, option space, $|\bar{O}|$ may grow to $|O|^{|\Upsilon|}$. The designer may use domain constraints that reduce $|\bar{O}|$, i.e. prevent some $\bar{o} \in O^1 \times \ldots \times O^{|\Upsilon|}$ from occurring.

We now define the subsets $O^j_d \subseteq O^j$ such that all options in O^j_d are specified at the same hierarchical level, d (of agent j); similarly we define $\bar{O}_d \subseteq O^1_d \times \ldots \times O^{|\Upsilon|}_d$ for $0 < d \leq$ *hierarchyDepth*, each constraining $\bar{O}$ to the options available at hierarchical level d (level-0 is the hierarchy root at which there are no options to choose from).

A policy over $\bar{O}_d$ is implicitly defined by the SMDP M_d which is defined over S with the *collective layer options* at level d as options. The M_d solution is the level-d constrained optimal meta-policy, i.e., it is the way to choose, at each state, the level-d individual policies which, in the long run, gather the highest collective reward.

Figure 1 illustrates the two-layer decision model where the individual layer (each agent^i task hierarchy) has 3 levels and thus the collective layer (represented by two, $\bar{o}_1$ and $\bar{o}_2$, possible multi-option instances) contains 2 levels; at each level, the set of diamond ended arcs, links the option $o_{\bar{o}}$ to its current policy $\pi_{\bar{o}}$.

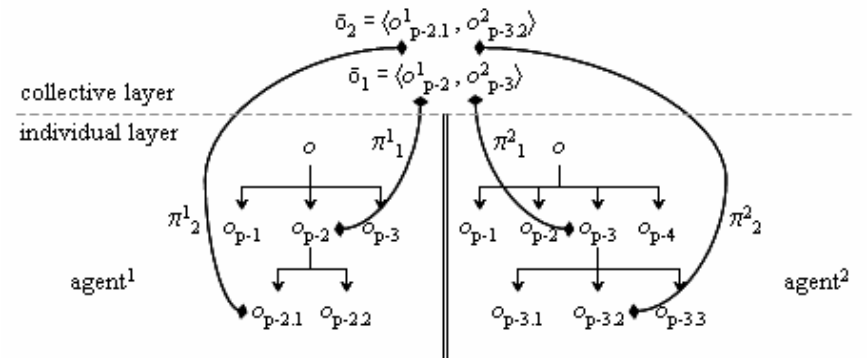

Fig. 1. The two-layer decision model and the links between layers (superscript j refers to agent^j; k subscript refers to the k hierarchical level and p-k subscript refers to the k tree path).

A centralized approach defines the $\mathcal{M}_d$ meta-policies and decides which individual policy to follow. Our approach is decentralized and lets each agent decide whether to make a decision by itself or to ask the collective layer for a decision.

In order to make well-informed decisions the collective layer needs to know, once an individual option terminates, about: i) the agent's world observation at that time instant, and ii) the agent's reward acquired since the last reported information. To exchange information we use the *request* and *inform* performatives' FIPA [6] pattern and we specified a simple protocol. At the individual layer: i) an option termination triggers an *inform* to the collective layer containing the agent's local observation and accumulated reward (since its last report), and ii) a *request* contains the option's hierarchal level at which the collective indication is required. At the collective layer: i) an *inform* arrival triggers a *request* for other agents to indicate their execution status, observation and cumulative reward (since last report), and ii) a *request* arrival triggers a decision about which admissible option the requesting agent should follow, *inform*'s the requester and assumes that the requester will follow that option's policy.

4 The CvI Regulatory Mechanism

The first step to instantiate a CvI model is to identify, from each agent's options set, $\mathcal{O}$, the subset of options, $\mathcal{C} \subseteq \mathcal{O}$, where a meta-policy is most effective to achieve coordination skills; the remaining options, $\mathcal{J} = \mathcal{O} - \mathcal{C}$, are purely individual. A simple, domain-independent design defines $\mathcal{C}$ as the set of multi-step options; hence $\mathcal{J}$ 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.

The $\mathcal{C}_d \subseteq \mathcal{C}$ and $\mathcal{J}_d \subseteq \mathcal{J}$ sets are used to define each $\bar{O}_d$ option space. At the collective layer, each option, $\bar{o}_d \in \bar{O}_d$, contains all $\mathcal{C}_d$ individual options, plus a special option, indOp_d, that represents $\mathcal{J}_d$ at the collective layer, i.e. indOp_d lets the agent itself choose among individual options. The indOp_d option is always admissible, terminates when the individual layer option being executed terminates and its policy is the agent's policy for choosing among purely individual options.

The second step is to devise the runtime mechanism that regulates the importance the agent credits to individual and collective decisions. The "importance" is a criterion related to a valuation in terms of benefits and costs an agent has of a mental state situation [4]; here, the mental state is the agent's policy space. We materialize the "importance" as the ratio between, the maximum expected benefit, in choosing a collective and an individual option. Thus, we define the *regulatory condition*: "$f(\max_c Q(s, c)) / f(\max_j Q(s, j)) < \kappa$", where $\kappa \geq 0$, $c \in \mathcal{C}_d$ and $j \in \mathcal{J}_d$ and $0 < f(x) = 1/(1+e^{-\chi x}) < 1$. The κ threshold is used to grade the focus from the individual to the collective layer. The f(x) normalizes the action-value function, Q, and the χ parameter configures the significance of "close" values (around the origin).

The usage of the regulatory condition (rC) depends on the kind of agent we want to implement. We used it to design agents that only consider the individual options that are expected to be better than any collective option, i.e. they follow the *behavior rule*: "`if` [rC $\wedge$ ($\max_j Q(s, j) > \max_c Q(s, c)$)] `then` decide-at-individual-layer, `else` *request*-collective-layer". Thus, the *behavior rule* defines: i) if $\kappa = 0$, then "always collective" decision, ii) if $\kappa \geq 1$, "when-possible", i.e. $\max_j Q(s, j) > \max_c Q(s, c)$, individual decision, and iii) if $0 < \kappa < 1$, decision depends on the rC value. The next sections show the cooperative behavior of these agents on different settings.

5 Experiment Specification

We implemented the CvI decision model and tested it in a multi-agent taxi problem (that extends the original single-agent taxi problem [5]), where a maze-like grid is inhabited by taxis (agents), passengers and sites. Passengers appear at a site and wish to be transported to another site. Taxis go to the origin site of a passenger, pick up the passenger, go to its destination site and drop down the passenger. Taxis may pick up several passengers; a site may have several passengers, each with its own destination.

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 is to learn a coordinated behavior where taxis cooperate to minimize the resources (time) spent to satisfy the passengers' needs.

We made 3 different specifications of individual and collective options: i) *CvI*, defines multi-step options as collective and one-step options as individual, ii) *purely collective*, only considers one-step options, all defined as collective, and iii) *purely individual*, considers that there are no collective options.

The same setup is used for all experiments: 5×5 grid, 4 sites $\mathbb{S}b = \{b1, b2, b3, b4\}$, 2 taxis $\mathbb{S}t = \{t1, t2\}$, and 2 passengers, psg_1 and psg_2. The actions available to each taxi are `pick`, `put`, and `move`(m), $m \in \{N,E,S,W\}$, the four cardinal directions.

The learning of the collective coordination policy occurs simultaneously with learning of the individual policies and the experiments' results (cf. Section 6) show that, except for the *purely individual*, all other agents exhibit a coordination policy. Also, each agent always learns optimal individual policies, i.e. the best way to execute tasks (e.g. how to navigate to a site and when to pick up a passenger) and their proper order execution (e.g. pick up a passenger at origin before navigating to destination).

Individual Layer Specification. The taxi observation, $\omega = \langle x, y, psg_1, psg_2 \rangle$, represents its own (x,y)-position and status of passenger, $psg_i = \langle loc_i, dest_i \rangle$, where $loc_i \in \mathbb{S}b \cup \mathbb{S}t \cup \{t1_{acc}, t2_{acc}\}$ (j_{acc} means taxi j accomplished delivery) and $dest_i \in \mathbb{S}b$. The rewards provided to a taxi are: i) 20 for delivering a passenger, ii) -10 for illegal `pick` or `put`, and iii) -1 for any other action, including moving into walls.

The same task hierarchy is used both at the *CvI* and at the *purely individual* specifications, which is composed of a `root` option and a `navigate`(b) option for each $b \in \mathbb{S}b$ (which, except for the root option is identical to the specification in [8]). The taxi's observation space is denoted by Ω, and $\Omega_b = \{\omega \in \Omega \mid x = x_b \wedge y = y_b\}$ represents the situations where the taxi is at site b. Thus, options are defined as: `navigate`(b) $\equiv \langle \mathbb{I}_b, \pi_b, \beta_b \rangle$, where i) $\mathbb{I}_b = \Omega - \Omega_b$, ii) π_b is the policy to learn, and iii) $\beta_b = 1$ if $b \in \Omega_b$ or $\beta_b = 0$ otherwise. The actions `move`(m), $m \in \{N,E,S,W\}$ are the only ones available to π_b policy. The root option of agent j, is `root` $\equiv \langle \mathbb{I}, \pi, \beta \rangle$, where i) $\mathbb{I} = \Omega$, and ii) $\beta = 1$ if ($\exists\ loc_i : loc_i = j_{acc} \wedge \neg\ \exists\ loc_i : loc_i = j$) or $\beta = 0$ otherwise, i.e. a taxi terminates an episode as soon as it delivers at least one passenger and there are no more passengers in that taxi. The options available to the π policy are: `navigate`(b) for each $b \in \mathbb{S}b$, `pick` and `put`.

Therefore, each agent holds an option hierarchy with 3 levels where `root` is the level-zero option, `navigate`(b), `pick` and `put` are the level-one options and `move`(m) are the level-two one-step options (defined for each `navigate`(b)).

Collective Layer Specification. The collective layer holds the agents' combined observation $\mathbb{s} = \langle ag^1, ag^2, psg_1, psg_2 \rangle$, where ag^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 *CvI* specification considers $\mathbb{C}$ = {`navigate`(b) for all $b \in \mathbb{S}b$} and $\mathbb{J}$ = {`pick`, `put`}. These are level-1 options, so $\mathbb{C}_1 = \mathbb{C} \cup \{\mathsf{indOp}_1\}$ and $\mathbb{J}_1 = \mathbb{J}$. The *purely collective* only defines one-step options, each being an option in $\mathbb{C}$. The *purely individual* considers none decision-making at the collective layer, thus defining $\mathbb{C} = \varnothing$.

Within this experimental toy world, an individual agent perceives 25,600 states, and the collective layer contains 640,000 states; a *purely individual* decision considers 6 options, while for *CvI* there are 25 collective options. Hence, the experiments

capture some complexity of a disaster response environment while learning a coordination policy.

6 Experiments and Results

Our experiments were conducted with two main purposes: i) measure the influence of the regulatory mechanism in the coordination learning process, and ii) compare the quality of the achieved coordination, regarding the agents' cooperation behavior.

We ran 4 experiments using the *CvI* specification, each with a different regulatory threshold (cf. Section 4). We ran an additional experiment to compare the *CvI*, *purely collective* and *purely individual* specifications. Each experiment executed 100 episodes. An episode always started in the same state and terminated as soon as all passengers reached their destination.

Policy learning used a temporal difference approach (SMDP Q-learning [13], [3]) with an ε-greedy exploration strategy, which picks a random action with probability ε and behaves greedy otherwise (i.e. picks the action with the highest estimated action value). Each experiment started with $\varepsilon = 0.15$ and ε always decayed 0.004 after each 3 episodes for the last-third of the experiment.

Influence of the Regulatory Mechanism. Figures 2 and 3 show the result of using the regulatory mechanism, throughout the learning process, with the *CvI* specification; each plotted line is labeled `c_`κ after the threshold, κ, used. Figure 2 shows the evolution of the "always collective", `c_0`, and "when-possible individual", `c_1`, (cf. Section 4) learning processes; Fig. 3 uses the collective layer's cumulative reward (at an experiment) and compares the performance for $\kappa \in \{1, 0.7, 0.3, 0\}$.

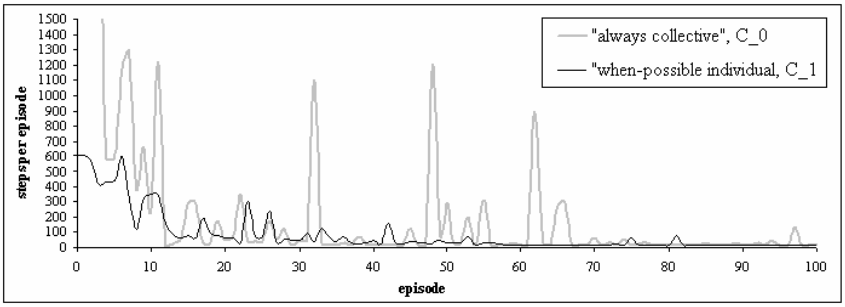

Fig. 2. The "always collective", `c_0`, and "when-possible individual", `c_1`, decisions.

Figure 2 shows that `c_0` is a slower and a more instable process than `c_1`. The higher complexity of the collective decision, `c_0`, is revealed mostly during the first and second thirds of the experiment. During the first-third of the experiment, the average number of steps per episode at `c_0` and `c_1` is, respectively 484 and 200; hence, an episode at `c_0` is, on the average, about 2.4 slower than an episode at `c_1`. The standard deviation is 694 at `c_0` (1.4 times average) and 178 at `c_1` (0.9 times average), which accounts for the instability, that although higher at `c_0`, occurs at

both processes. This relative behavior occurs again during the second-third of the experiment, although with lower average and standard deviation values. During the last-third of the experiment both `c_0` and `c_1` show a similar convergent pattern.

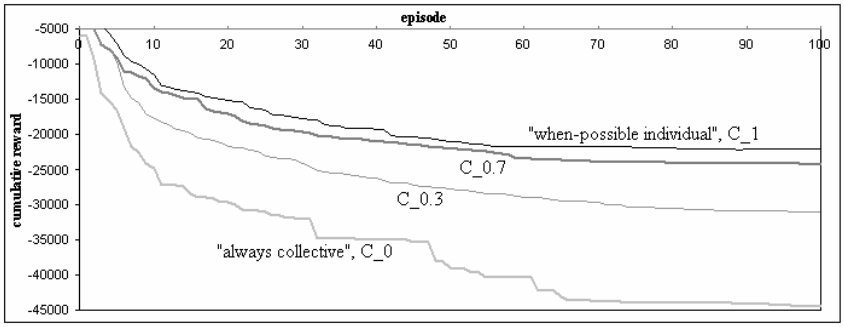

Fig. 3. The influence of the regulatory threshold in the "collective to individual tuning".

Figure 3 confirms that the influence of the regulatory mechanism is most relevant for the first-third of the experiment, where lower values of κ yield a faster decrease of cumulative rewards. It clearly shows that as we move from individual to collective focus (κ decreases) learning a coordination policy becomes more expensive.

All plotted experiences eventually converge to a performance (slope) that, although similar, represents distinct coordination policies (cf. Fig. 5).

An insight on these results is that as κ increases the agent is "left alone" and begins earlier to learn its task; then, when after some episodes the collective options become determinant to solve the task (higher action-values), the opportunity arises for the collective layer to choose options thus "compensating" the initial individual learning.

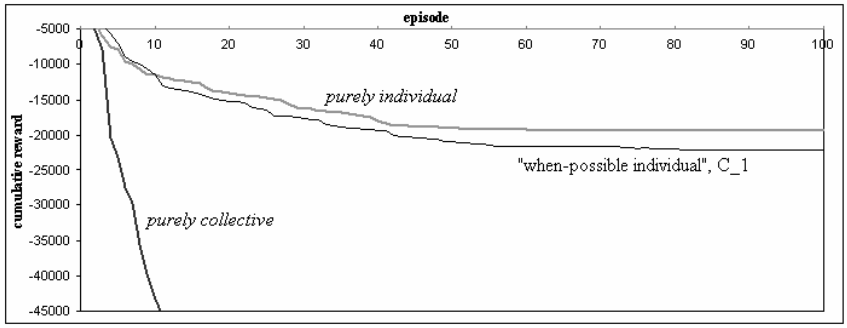

Fig. 4. The relation among specifications: *purely collective*, *purely individual* and *CvI* for κ=1.

We recall that, in the "always-collective" case, each agent owns a hierarchical task decomposition and the collective layer may tell an agent to execute an individual option (via `indOp`) which gives the agent some freedom to decide what to do.

In the *purely collective* case, the collective layer always tells an agent what to do. Figure 4 *purely collective* graph shows a high slope that rapidly decreases rewards, meaning that the collective layer is quickly getting lost in the confusion of low-level

details while trying to coordinate the agents (rewards reach ≈ -500,000, at the end of the experiment, still decreasing).

Hierarchical approaches, *purely individual* and `c_1`, exhibit a similar performance with `c_1` being asymptotically better since it leads to higher expected reward; `c_1` coordination policy takes 17 steps when *purely individual* needs 21 steps (cf. Fig. 5).

The Coordination Policies. The performance analysis does not evidence the concrete coordination policy learnt by each process. In order to conclude about coordination we need to analyze the details of a specific episode.

Figure 5 compares the coordination policies achieved at the last episode of `c_0`, `c_0.3`, `c_0.7`, `c_1` and *purely individual*. Each grid shows a different coordination policy, used by the taxis t_1 and t_2, to deliver passengers psg_1 and psg_2 respectively to the s_2 and s_1 sites.

Figure 5 (a) exhibits the optimal coordination strategy, where taxi t_1 although closer to psg_1 decides to pick up psg_2 and deliver at s_1; meanwhile t_2 picks up and delivers psg_1 at s_2. This strategy only takes 16 time steps to terminate the episode because taxi t_1 learnt to follow the cooperative collective perspective.

Figure 5 (b) coordination strategy takes 17 time steps to terminate because t_1 takes the individual perspective and decides to immediately pick up passenger psg_1; taxi t_2 cooperates with t_1's decision as t_2 decides to pick up psg_2 instead of running for psg_1.

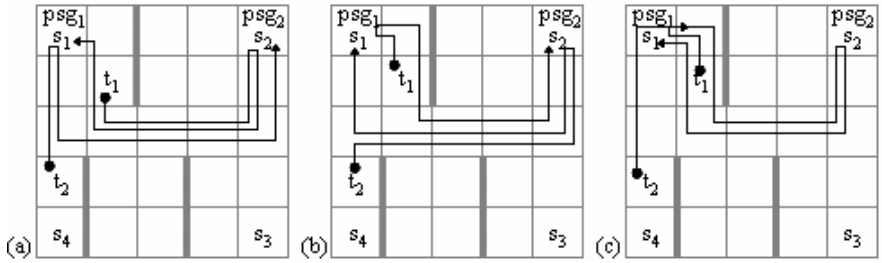

Fig. 5. The coordination policies, achieved by: (a) `c_0` and `c_0.3` with 16 steps, (b) `c_0.7` and `C_1` with 17 steps, and (c) *purely individual* with 21 steps.

Figure 5 (c) shows the individual perspective of taxis t_1 and t_2 as they both decide to pick up their closest passenger and they end up competing to pick up the same, psg_1, passenger. Taxi t_1 wins and picks up psg_1 after which they compete again to pick up psg_2. They both arrive at the same time at s_2 and it is interesting to notice that t_1 wins again and picks up psg_2. Taxi t_1 wins because instead of immediately dropping down psg_1 (that t_1 was carrying) it first picks up psg_2 and only then it drops down psg_1.

7 Conclusions

This work explores the separation of concerns between collective and individual decisions while learning coordination policies in a multi-agent complex environment. The separation of concerns frames a two-layer decision model where the definition of inter-layer relations enables the agent to decide at which behavioral layer a decision

should be taken. Such two-layered decision-making represents our novel contribution to multi-agent coordination within a reinforcement learning framework.

The formulation of the two-layer, CvI, multilevel hierarchical decision model and the definition of an inter-layer regulatory mechanism enable us to show experimentally how to explore the individual policy space in order to decrease the complexity of learning a coordination policy in a partially observable setting.

The two-layer approach augments the design flexibility in two ways: i) makes it possible to specify individual task hierarchies that are not necessarily equal, therefore allowing for agents' heterogeneity, and ii) enables to configure different architectures (e.g. centralized or decentralized) depending on the information exchange between collective and individual layers.

Future work will explore, at the collective layer, the inclusion of state abstraction [8] and the integration of teamwork strategies [14] to search the most distinguishing state space and to reduce the space of admissible actions, respectively. We aim to apply the resulting technique to a simulated disaster response environment [9].

Acknowledgements. Research was partially supported by PRODEP III 5.3/13/03.

8 References

[1] Abdallah, S. and Lesser, V. (2005). Modeling Task Allocation Using a Decision Theoretic Model. Fourth Int. Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS05), pp. 719–726. ACM Press.

[2] Boutilier, C. (1999). Sequential Optimality and Coordination in Multi-Agent Systems. Sixteenth Int. Joint Conference on Artificial Intelligence (IJCAI99). pp. 478–485.

[3] Bradtke, S. and Duff, M. (1995). Reinforcement learning methods for continuous time Markov decision problems. Advances in Neural Inf. Processing Systems 8, pp. 393–400.

[4] Corrêa, M. and Coelho, H. (2004) Collective Mental States in Extended Mental States Framework. International Conference on Collective Intentionality.

[5] Dietterich, T. (2000). Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition. Artificial Intelligence Research 13, pp. 227–303.

[6] FIPA Communicative Act Library Specification, http://www.fipa.org, 2002.

[7] Ghavamzadeh, M.; Mahadevan, S.; Makar, R. (2006). Hierarchical Multi-Agent Reinforcement Learning. Journal of Autonomous Agents and Multi-Agent Systems.

[8] Jonsson, A. and Barto, A. (2001). Automated State Abstractions for Options Using the U-Tree Algorithm. Advances in Neural Inf. Processing Systems, 13, pp. 1054–1060.

[9] Kitano, H.; Tadokoro, S.; Noda, I.; Matsubara, H.; Takahashi, T.; Shinjou A.; Shimada, S. (1999). RoboCup Rescue: Search and Rescue in Large-Scale Disasters as a Domain for Autonomous Agents Research. Conf. on Man, System and Cyb. (MSC-99), pp. 739–743.

[10] Nash, J. (1951). Non-Cooperative Games. Annals of Mathematics, 54, pp. 286-295.

[11] Pynadath, D. and Tambe, M. (2002). The Communicative Multiagent Team Decision Problem: Analyzing Teamwork Theories and Models. Journal of AI Research, 389–423.

[12] Rohanimanesh, K. and Mahadevan, S. 2003. Learning to Take Concurrent Actions. Sixteenth Annual Conference on Neural Information Processing Systems, pp 1619–1626.

[13] Sutton, R.; Precup, D.; and Singh, S. (1999). Between MDPs and Semi-MDPs: A framework for temporal abstraction in reinforcement learning. A.I., 112(1-2), pp181–211.

[14] Trigo, P. and Coelho, H. (2005). The Multi-Team Formation Precursor of Teamwork. EPIA05, LNAI 3808, pp. 560–571. Springer-Verlag.