Multiagent Metareasoning Through Organizational Design

Jason Sleight and Edmund H. Durfee
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109
{jsleight,durfee}@umich.edu

Abstract

We formulate an approach to multiagent metareasoning that uses organizational design to focus each agent's reasoning on the aspects of its local problem that let it make the most worthwhile contributions to joint behavior. By employing the decentralized Markov decision process framework, we characterize an organizational design problem that explicitly considers the quantitative impact that a design has on both the quality of the agents' behaviors and their reasoning costs. We describe an automated organizational design process that can approximately solve our organizational design problem via incremental search, and present techniques that efficiently estimate the incremental impact of a candidate organizational influence. Our empirical evaluation confirms that our process generates organizational designs that impart a desired metareasoning regime upon the agents.

1 Introduction

When autonomous agents operate in large, complex, and time-critical problem domains, the amount of computation time needed to make provably optimal decisions can exceed the time available before action must be taken. Research into metareasoning—reasoning about reasoning—studies mechanisms that agents can use to make principled decisions about whether the improvements to decisions from additional reasoning are expected to outweigh the costs of delaying enacting decisions. (See Cox and Raja (2011) for a thorough discussion of work in this area.) Metareasoning becomes even more complicated in multiagent settings, since the benefits of additional reasoning might depend on the reasoning and behaviors of other agents (Raja and Lesser 2007). For example, if one agent assumes responsibility for (reasoning about) performing a task, then there might be no additional benefit for other agents to also reason about that task. Thus, research into multiagent metareasoning has been formulated as a metacoordination problem, where agents individually make metareasoning decisions but coordinate those decisions to strike a good collective balance between their expected joint performance and reasoning costs (Raja and Lesser 2007; Alexander et al. 2007).

In this paper, we investigate an alternative approach to solve the multiagent metareasoning problem through organizational design, where a good multiagent organization should both guide agents into coordinated behaviors, and also guide them into coordinated reasoning about their individual decision problems. Of course, this approach simplifies the multiagent metareasoning problem that the agents face by complicating the organizational design problem to find a design that not only leads to coordinated behavior in the world, but also coordinated utilization of the agents' distributed reasoning resources. In the first contribution of this paper (Section 3), we leverage the decentralized Markov decision process (Dec-MDP) formalism to characterize an organizational design problem that explicitly considers the quantitative impact that an organizational design has on both the expected performance of the agents' behaviors as well as the reasoning demands placed on the agents.

In Section 4, we make our second contribution by formulating an automated organizational design process (ODP) to solve our organizational design problem. Unsurprisingly, we find that creating an optimal organizational design is computationally intractable, and thus we develop techniques to improve the ODP's computational efficiency at the expense of optimality guarantees. Namely, we describe methods for efficiently estimating the incremental impact of an individual organizational influence, and illustrate how an ODP can embed these calculations within an incremental search of the organizational influence space. In Section 5, we empirically evaluate our organizational design algorithm, and find that our ODP finds good organizational designs that impart a target metareasoning regime upon the agents. We end the paper (Sections 6 and 7) by discussing how our work relates to other research and briefly summarizing our results. We next (Section 2) more formally describe our agents' reasoning framework and a simplified firefighting domain that we use for illustration and experimentation throughout this paper.

2 Problem Domain

We assume a multiagent system consisting of n fully-cooperative decision-theoretic agents, where each agent i begins with its own local decision-theoretic model, $\mathcal{M}_i$, that captures its local view of the true, global decision-theoretic environment model, $\mathcal{M}^*$. For this paper, we assume the $\mathcal{M}_i$s compose a locally-fully observable Dec-MDP (Becker et al. 2004), and thus $\mathcal{M}_i = \langle S_i, \alpha_i, A_i, P_i, R_i, T_i \rangle$ which specifies agent i's local state space (S_i), initial state distribution

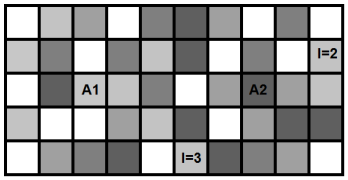


Figure 1: Example initial state in the firefighting grid world. Ai is the position of agent i, and $I = x$ indicates a fire in that cell with intensity x. Darker shading indicates higher delay.

(α_i), action space (A_i), transition function (P_i), reward function (R_i), and finite time horizon (T_i). The various agents' $\mathcal{M}_i$s need not be independent, and there may be overlapping information (e.g., state factors) among the agents. Each agent can use its $\mathcal{M}_i$ to calculate an optimal local policy $\pi_i^* : S_i \times A_i \mapsto [0, 1]$ via any of the standard policy creation techniques (Puterman 1994). The joint policy is defined as $\pi = \langle \pi_1^*, \cdots, \pi_n^* \rangle$, which may or may not be jointly optimal depending on how the agents coordinate the construction of their local policies. The Q-value, $Q^\pi(s, a)$, is the expected cumulative reward of executing joint action a in global state s and then following joint policy π.

To illustrate a problem of this type, we reuse a simplified firefighting scenario (Sleight and Durfee 2013), where firefighting agents and fires to be fought are in a simulated grid world. The global state consists of: a system time; the locations of the agents; the fire intensity, $I_c \in \mathbb{Z}^+$, for each cell c; and a delay, $\delta_c \in [0, 1]$, for each cell c. Figure 1 shows an initial global state for a two-agent problem, where the locations of agents A1 and A2 are shown, along with the intensities of fires in the two cells with $I_c > 0$, and darker shaded cells have higher delay. An agent does not observe the position of the other agent, but otherwise can observe the global state. Each agent has 6 local actions: a NOOP action that only increments system time (as do all other actions); 4 possible movement actions (N, S, E, W) that move the agent one cell in the specified direction (into cell c) with probability $1 - \delta_c$, and otherwise behaves like a NOOP; and a fight-fire (FF) action that decrements by 1 the intensity of the agent's current cell (to a minimum of 0) and otherwise behaves like a NOOP. Joint actions are defined as the combination of the agents' local actions. Movement actions are independent (agents can occupy the same location), but FF actions are not: the intensity of a cell only decreases by 1 even if multiple agents simultaneously fight a fire. The local reward for each agent in each state is the negative sum of fire intensities in that state, and agents plan for a fixed time horizon.

We also assume a temporally-extended environment, where the agents will solve a sequence of distinct problem episodes, and the environment returns to a fixed, recognizable state between episodes. For example, in the firefighting domain, in each episode the agents fight a set of fires and between episodes return to a state with no fires and where agents are back at their initial positions. Though the episodes are distinct, the benefit of a long-term organizational design relies on an ODP's ability to identify and codify repeated patterns of reasoning and interactions within and across the episodes. For example, though fires usually are located in different cells each episode, an ODP might find patterns over which regions each agent should typically be responsible for and codify those patterns as organizational influences. This type of episodic behavior is prevalent in a wide range of domains such as emergency response, distributed sensor networks, supply chain management, and most traditional long-term organizations (e.g., a university or business).

3 Metareasoning Via Organizational Design

The idea that organizational designs can impact agents' reasoning and behaviors is well-established. For example, social laws (Shoham and Tennenholtz 1995) affect the reasoning that agents perform as well as the behaviors they execute. To our knowledge, however, no prior work has *explicitly* leveraged this capability within an ODP to *intentionally* impart a specific, desired metareasoning regime upon the agents, i.e., a specific tradeoff between the agents' reasoning costs and performance of their behaviors. In this section, we formulate an organizational design problem that quantitatively incorporates both the expected performance of the agents' behaviors as well as their expected reasoning costs to achieve those behaviors. In contrast to typical metareasoning approaches which try to dynamically assess the predicted benefit of additional reasoning (Hansen and Zilberstein 2001b), the fundamental idea of our approach is to have an ODP utilize its global view of the problem domain to identify high-performing behavior patterns, and then influence the agents to avoid *even thinking* about behaving counter to those patterns. For example, using its global perspective, an ODP might identify that agents should typically fight fires near their initial locations. It might then codify this pattern by restricting an agent from reasoning about fighting fires in distant cells, which imposes a metareasoning regime that trades computational speedup (due to never considering fighting fires in those distant cells) for small expected reward loss (in the rare cases that it should fight those fires).

We define an **organizational influence**, Δ_i, as a modification to agent i's local problem description, $\mathcal{M}_i$, that either constrains the agent's local policy space, or re-prioritizes the agent's preferential ordering over its local policy space. An **organizational design**, Θ, is a set of organizational influences for each agent, $\Theta \equiv \langle \theta_1, \cdots, \theta_n \rangle$, where $\theta_i \equiv \{\Delta_i\}$ is the set of organizational influences for agent i. Let $\pi^{|\Theta} = \langle \pi_1^{*|\theta_1}, \cdots, \pi_n^{*|\theta_n} \rangle$ refer to the agents' joint policy w.r.t. Θ, where $\pi_i^{*|\theta_i}$ refers to agent i's optimal local policy w.r.t. θ_i.

Leveraging Dec-MDP principles, we can quantitatively measure the performance of an organizational design Θ. The **operational reward** under Θ, $\mathbb{R}_{Op}(\Theta)$, is given by the expected joint reward of $\pi^{|\Theta}$:

$$\mathbb{R}_{Op}(\Theta) \equiv \sum_{s \in S} \alpha(s) \sum_{a \in A} \pi^{|\Theta}(s, a) Q^{\pi^{|\Theta}}(s, a)$$

Assuming agents reason in parallel, the **operational reasoning cost** under Θ, $\mathbb{C}_{Op}(\Theta)$, is given by the expected operational reasoning cost for an agent to calculate its individual

$\pi_i^{*|\theta_i}$, notated as $C(\pi_i^{*|\theta_i})$:

$$\mathbb{C}_{Op}(\Theta) \equiv E_i[C(\pi_i^{*|\theta_i})]$$

To measure the quality of the metareasoning regime imparted by Θ, we combine these metrics in the **operational performance** of Θ, $\mathbb{P}_{Op}(\Theta)$, which is a function, f, of the operational reward and reasoning cost under Θ.

$$\mathbb{P}_{Op}(\Theta) \equiv f(\mathbb{R}_{Op}(\Theta), \mathbb{C}_{Op}(\Theta)) \quad (1)$$

The specific form of f is defined by the problem domain, and f conveys information to the ODP so it can determine how it should trade off between $\mathbb{R}_{Op}(\Theta)$ and $\mathbb{C}_{Op}(\Theta)$.

The optimal organizational design is given by $\Theta^* \equiv \text{argmax}_\Theta \mathbb{P}_{Op}(\Theta)$. Unfortunately, the space of possible Θs is intractably large for even small, simple domains, making direct enumeration infeasible. By our earlier definition, an organizational design is a set of constraints and re-prioritizations over the agents' policy spaces, and thus there is at least one organizational design for each possible total ordering of every subspace of the joint policy space, yielding $|\boldsymbol{\Theta}| = \sum_{i=1}^{|\boldsymbol{\pi}|} \frac{|\boldsymbol{\pi}|!}{(|\boldsymbol{\pi}|-i)!} = O(|\boldsymbol{\pi}|!)$ as a lower bound for the worst-case complexity, where $|\boldsymbol{\pi}| = O(|A_i|^{|S_i|^n})$. In the next section, we describe several approximations an ODP can make to efficiently search through this space while still finding an organizational design that imparts a good (but not provably optimal) metareasoning regime upon the agents.

4 Organizational Design Process

We focus on incremental search, and on techniques for computing the incremental impact of an individual influence, since as shown in Section 3 direct enumeration of the organizational design space is infeasible. We utilize a simple greedy hill-climbing search, although other incremental search algorithms (e.g., Monte Carlo, A*, etc.) could be used instead. Characterizing the performance profiles of alternative incremental search algorithms as a function of a problem domain's operational performance topology is something we plan to consider in the future.

Naïvely, a greedy algorithm computes iteration $j+1$ as:

$$\Theta^{j+1} = \Theta^j + \underset{\Delta_i, \forall i \in n}{\text{argmax}} \mathbb{P}_{Op}(\Theta^j + \Delta_i) \quad (2)$$

where $\Theta^j + \Delta_i \equiv \langle \theta_1^j, \cdots, \theta_i^j \cup \Delta_i, \cdots, \theta_n^j \rangle$. Notice however, that Equation 2 requires recomputing the performance contribution of Θ^j for each $\mathbb{P}_{Op}(\Theta^j + \Delta_i)$, which could waste substantial computational effort. If we can instead factor the calculation of $\mathbb{P}_{Op}(\Theta^j + \Delta_i)$ into $\mathbb{P}_{Op}(\Theta^j)$ and the conditional, incremental impact of Δ_i w.r.t. Θ^j, then we could avoid this redundant computation. We achieve this by linearly approximating $\mathbb{P}_{Op}(\Theta^j + \Delta_i)$. Assuming $\mathbb{R}_{Op}$, $\mathbb{C}_{Op}$, and $\mathbb{P}_{Op}$ are everywhere differentiable[1], and abusing notation to write the linear approximations in traditional form, we get:

$$\mathbb{R}_{Op}(\Theta^j + \Delta_i) \approx \mathbb{R}_{Op}(\Theta^j) + \Delta_i \cdot \frac{d\mathbb{R}_{Op}}{d\Theta^j}(\Theta^j)$$

$$\mathbb{C}_{Op}(\Theta^j + \Delta_i) \approx \mathbb{C}_{Op}(\Theta^j) + \Delta_i \cdot \frac{d\mathbb{C}_{Op}}{d\Theta^j}(\Theta^j)$$

[1]While the everywhere differentiable assumptions are theoretically required, in practice we have not found them necessary.

$$\mathbb{P}_{Op}(\Theta^j + \Delta_i) \approx f(\mathbb{R}_{Op}(\Theta^j), \mathbb{C}_{Op}(\Theta^j)) + \Delta_i \cdot \frac{d\mathbb{R}_{Op}}{d\Theta^j}(\Theta^j) \frac{\delta f}{\delta \mathbb{R}_{Op}}(\Theta^j) + \Delta_i \cdot \frac{d\mathbb{C}_{Op}}{d\Theta^j}(\Theta^j) \frac{\delta f}{\delta \mathbb{C}_{Op}}(\Theta^j)$$

Substituting the above equations into Equation 2 yields:

$$\Theta^{j+1} = \Theta^j + \underset{\Delta_i}{\text{argmax}} \left[\Delta_i \cdot \frac{d\mathbb{R}_{Op}}{d\Theta^j}(\Theta^j) \frac{\delta f}{\delta \mathbb{R}_{Op}}(\Theta^j) + \Delta_i \cdot \frac{d\mathbb{C}_{Op}}{d\Theta^j}(\Theta^j) \frac{\delta f}{\delta \mathbb{C}_{Op}}(\Theta^j) \right] \quad (3)$$

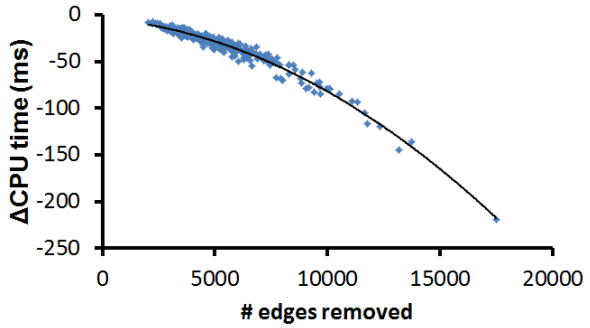
which avoids redundantly computing how Θ^j impacts the operational performance, and instead only computes the incremental impact of Δ_i on operational performance.

We previously (Sleight and Durfee 2013) examined how to express organizational influences to decision-theoretic agents such that each influence has well-defined impact on the agents' reasoning processes, which we leverage here to delineate the space of influences an ODP can consider. In our prior work, we found that representing influences as modifications to factors of the agents' local decision problems (i.e., $S_i, \alpha_i, A_i, P_i, R_i$, and T_i) provides an expressive specification language that the agents can easily incorporate into their planning processes. In Sections 4.1 and 4.2, we describe a general methodology for efficiently computing $\Delta_i \cdot \frac{d\mathbb{C}_{Op}}{d\Theta^j}(\Theta^j)$ and $\Delta_i \cdot \frac{d\mathbb{R}_{Op}}{d\Theta^j}(\Theta^j)$ respectively for any of our previously identified influence forms, and then in Section 4.3 we illustrate in detail how an ODP can implement this methodology for action influences.

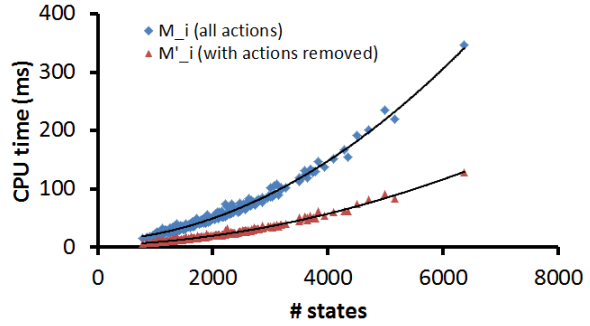
4.1 Computing Incremental Reasoning Costs

$\Delta_i \cdot \frac{d\mathbb{C}_{Op}}{d\Theta^j}(\Theta^j)$ corresponds to the conditional impact to the agents' computational costs from adding Δ_i w.r.t. Θ^j. An agent's computational costs are determined by two primary factors (Littman, Dean, and Kaelbling 1995), the number of states in its planning problem and the number of edges in its state graph. Thus, determining incremental reasoning costs relies on determining the expected marginal costs of adding a new state/edge, and then calculating the expected change to the number of states and edges caused by adding Δ_i into Θ^j.

Our methodology for empirically estimating the marginal cost of a state and/or edge is as follows. We have an agent first use its local model $\mathcal{M}_i$ to compute π_i^* in a set of episodes. We then create a modified version of $\mathcal{M}_i$, labeled as $\mathcal{M}_i'$, that contains the minimal number of edges between states such that the reachable state space from all possible initial states is unchanged, and the optimal policy is unchanged. We include the latter condition so that the bias of our estimate matches desired ODP behavior of removing non-optimal behaviors. The agent then solves the problem set again, but plans using $\mathcal{M}_i'$ instead of $\mathcal{M}_i$. Taking the relative computational difference between these experiments provides an empirical estimate of an edge's marginal cost. Additionally, since $\mathcal{M}_i'$ is "minimally" connected, the relative computational difference across the episodes (which typically have different numbers of states), provides a good estimate of a state's marginal cost, i.e., "maximally" disentangles the cost of a state and the edges to connect it to the state space.



(a) CPU savings vs. # of edges removed from agent's problem



(b) CPU time for agent vs. # of states in agent's problem

Figure 2

To demonstrate the use of this methodology, Figure 2 shows its application in the firefighting domain, using 300 randomly-generated episodes. For reference, $\mathcal{M}'_i$ removes approximately 2.6 edges for every state. Taking the derivative of Figure 2a shows that an edge's marginal computational cost is approximately $1.2e_i + 2000$ (ns), where e_i is the current number of edges. Taking the derivative of the $\mathcal{M}'_i$ line in Figure 2b shows that a state's marginal cost is approximately $5.28s_i + 3000$ (ns), where s_i is the current number of states. The exact values we found here are clearly only applicable for our agents' specific policy creation implementation within the firefighting domain; however, our methodology generalizes to any problem domain expressed as a Dec-MDP, and to any Dec-MDP solution techniques.

4.2 Computing Incremental Reward

$\Delta_i \cdot \frac{d\mathbb{R}_{Op}}{d\Theta^j}(\Theta^j)$ corresponds to the expected Q-value change from adding Δ_i into Θ^j. By definition, $Q^\pi(s,a)$ only changes if Δ_i alters either π, the immediate reward $R(s,a)$, or the transition probabilities $P(s'|s,a)$. Since alterations to $R(s,a)$ or $P(s'|s,a)$ also induce changes to π (and otherwise Δ_i's impact on Q-values is trivial to compute), we focus on how Δ_i alters the agents' policy w.r.t. Θ^j. While the ODP could do this by calculating $\pi^{|\Theta^j+\Delta_i}$ for each candidate organizational design, such an approach is computationally daunting given the complexity of computing optimal policies and the possible number of candidates. Instead, the insight we exploit is that an ODP can use its global view to compute/estimate an optimal joint policy, π^*, once, and then should only consider candidate organizations that preserve this policy while steering agents away from taking, and even considering, behaviors outside of this policy. If the organization does not preclude π^*, then the calculation of Δ_i's impact to the agents' policy is independent of Θ^j, and the ODP does not need to compute $\pi^{|\Theta^j+\Delta_i}$. While the ODP (unavoidably) must still determine what good behaviors are by calculating an optimal joint policy, the ODP only need do this costly calculation once—rather than $O(|\Delta_i|^2)$ times—and then amortize those costs over all of the search iterations, which results in substantial computational savings.

4.3 Action Influences

In this section, we illustrate how an ODP can apply our general methodology from Sections 4.1 and 4.2 to action influences. We chose to implement action influences because they are a particularly commonplace organizational mechanism in previous research (Shoham and Tennenholtz 1995; Pacheco and Carmo 2003; Horling and Lesser 2008). Note however, that prior work has not given explicit, quantitative consideration to how such influences affect the agents' metareasoning regime, which is our focus here.

An action influence, Δ_i, that blocks action a_i from consideration in state s_i, removes one edge for each possible successor state upon taking a_i in s_i, and removes any now-unreachable states. By enumerating the successor states (via the transition function), an ODP can calculate the expected change to the number of edges, $|E_i^{\Delta_i}|$, and states $|S_i^{\Delta_i}|$, caused by adding Δ_i to Θ^j. Combining those quantities with our previous marginal cost estimates in Section 4.1 yields:

$$\Delta_i \cdot \frac{d\mathbb{C}_{Op}}{d\Theta^j}(\Theta^j) = \left(5.28|S_i^{\Theta^j}| + 3000\right)|S_i^{\Delta_i}| + \left(1.2|E_i^{\Theta^j}| + 2000\right)|E_i^{\Delta_i}|$$

where $|S_i^{\Theta^j}|$ and $|E_i^{\Theta^j}|$ are the expected number of states and edges respectively for agent i given that it conforms to Θ^j. $|S_i^{\Theta^j}|$ and $|E_i^{\Theta^j}|$ are known from the previous search iteration, meaning this computation requires only $O(|S_i^{successor}|)$ time for enumerating the successor state space.

The expected Q-value change associated with an action influence Δ_i, that blocks action a_i from consideration in state s_i, is equal to the expected difference between the Q-value of a_i and the next best action. Mathematically this yields,

$$\Delta_i \cdot \frac{d\mathbb{R}_{Op}}{d\Theta^j}(\Theta^j) = E_{s \mapsto s_i}\left[\left(\left(\max_{a=\langle\cdot,a_i,\cdot\rangle} Q^{\pi^*}(s,a)\right) - \left(\max_{a'\neq\langle\cdot,a_i,\cdot\rangle} Q^{\pi^*}(s,a')\right)\right)x(s,a)\right]$$

where (in a non-recurrent state space like the domains we consider here), an occupancy measure, $x(s,a)$, is equal to the probability of reaching state s and then executing action a, and is calculated via a dual problem representation (Kallenberg 1983). This computation requires $O(|A||S|)$ time in the worst case, but the $|S|$ term represents the number of states that map into s_i and will often be much less than the total number of global states.

5 Evaluation

We begin our evaluation by briefly describing some implementation details of our evaluation domain, and ODP. Our experiments use the firefighting domain as previously described in Section 2, where in each episode there are: two agents, who always begin in the initial locations in Figure 1; two fires, each with initial intensity independently and uniformly selected from $\{1, 2, 3\}$, and with a uniformly random, but distinct location; delay in each cell independently and uniformly chosen from $[0, 1]$; and a time horizon of 10. Agents create their optimal local policies with respect to their organizationally augmented local model using CPLEX (IBM 2012) to solve a linear program (Kallenberg 1983). While the agents' individual planning problems in this domain are relatively simple, it is important to recognize that the difficulty of the ODP's problem instead stems from the complexity of the possible agent-interaction patterns, which given the breadth of ways agents' local models can interact, makes the firefighting domain interesting from an ODP perspective.

Previous research has shown that abstract organizational influences outperform detailed organizational micromanaging when agents possess local expertise (Dignum, Vázquez-Salceda, and Dignum 2005; Sleight and Durfee 2013). We incorporate this principle in two ways: first by presenting our ODP with a model where it only knows the mean cell delay as opposed to the specific delay of each cell for an episode, and second by having our ODP consider action influences for each agent i, that block an action, a_i, from an abstract local state, $\hat{s}_i$, where the abstraction drops all state factors excluding agent i's position. This abstraction was chosen to prevent the ODP from micromanaging the agents because it forces an influence to apply to broader situations. Finer abstractions would enable the ODP to find more refined organizational designs at the expense of greater ODP computation and/or overfitting (and *vice versa* for coarser abstractions).

Our ODP sampled and solved training problems from its domain model until it had stable estimates for $\Delta_i \cdot \frac{d\mathbb{R}_{Op}}{d\Theta^j}(\Theta^j)$, which took 300 samples in our experiments. To test our claim that our algorithm correctly finds an organizational design that imparts a desired metareasoning regime upon the agents, we explored a space of environments with a range of metareasoning tradeoff demands, parameterized by $\mathbb{P}_{Op}(\Theta) = \mathbb{R}_{Op}(\Theta) - \mathbb{C}_{Op}(\Theta)/b$ for different values of b. We present results across b values such that at extremely costly reasoning ($b = 1\text{E}4$) the ODP designs an organization where the agents only consider executing a single action (FF in this case), and at extremely low reasoning costs ($b = 1\text{E}8$) designs an organization where every action the ODP expects an agent to ever want to execute is included. Note that the latter, 1E8Org will still exclude local actions that would never be sensible (e.g., fighting fires in distant cells that are always another agent's responsibility).

Unexpectedly, we found that our ODP was able to encode surprisingly nuanced organizational designs despite being limited to a space of abstracted influences. For example, the ODP frequently imposes unidirectional movements (see Figure 3), where an agent is allowed to consider moving into a cell, but the action to move back and in effect "undo" the previous action is blocked from consideration. This type of influence imparts a good metareasoning regime by forcing the agent to reason about complete, irreversible behavior trajectories rather than needlessly reasoning about reversing prior actions. These unidirectional movements also improve coordinated behavior by discouraging an agent from rushing to the other side of the grid (where the other agent is located) to fight a high-intensity fire since it would be unable to come back and fight an initially-closer fire. In the future, we plan to investigate whether irreversible task trajectories are an effective general-purpose organizational strategy (especially with respect to metareasoning issues), as well as the possibility of other overarching heuristic organizational principles.

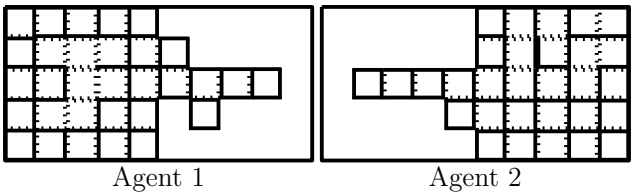
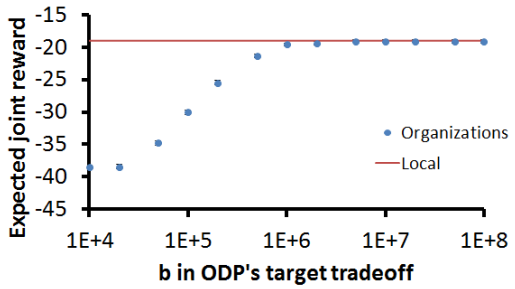
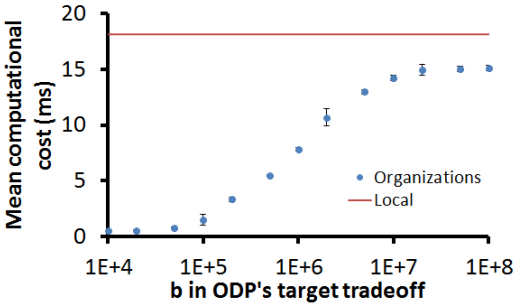


Figure 3: Agents' movement action influences in the bOrg with $b = 1\text{E}6$. An agent can move into a cell in a direction where it first passes a dotted line, but not a solid line.

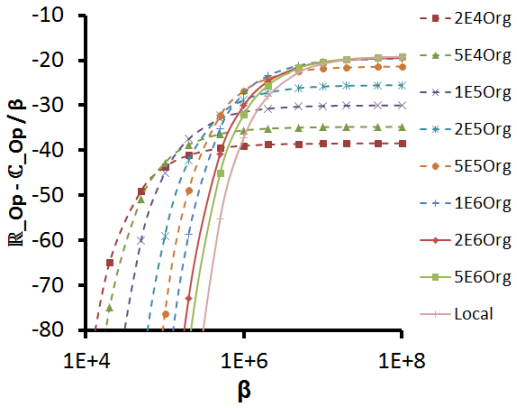
To quantitatively determine the expected joint reward and agent computational cost characteristics of each organizational design, we gave the agents a sequence of 1500 test problem episodes randomly sampled from $\mathcal{M}^*$ and had them utilize each of the organizational designs (as well as a local baseline with no organizational influences) in each of the episodes. We focus on mean performance over all problems not only to smooth out the inherent randomness of the episodes but also to emphasize that organizations are designed for long-term use over an extended time frame. Figures 4a and 4b show the mean $\mathbb{R}_{Op}$ and $\mathbb{C}_{Op}$ respectively over the 1500 test episodes for each of the bOrgs and the local baseline. These graphs show that as our organizational design algorithm faces different target metareasoning tradeoffs (i.e., values of b), the organizational designs it creates have monotonically increasing performance properties in both $\mathbb{R}_{Op}$ and $\mathbb{C}_{Op}$. That is, as computation becomes cheaper (b increases), the algorithm creates organizational designs that induce the agents to consider more actions (and thus utilize more computation), which yields increased expected joint reward. We also observe that these bOrgs, which are limited to influences that only remove actions from consideration, do not lead to agents finding better policies than they otherwise would have ($\mathbb{R}_{Op}$ of the bOrgs do not surpass the local baseline), but find these policies with significantly less computation (lower $\mathbb{C}_{Op}$). In Figure 4c, we use the expected $\mathbb{R}_{Op}$ and $\mathbb{C}_{Op}$ data to calculate the metareasoning regime imparted upon the agents by the organizations as a function of tradeoff parameterizations. This graph shows that, for any target tradeoff parameterization β, the best organizational design (i.e., maximizing the y-axis) is approximately the bOrg our algorithm generates with $b = \beta$, which confirms that our ODP designs organizations that approximately optimize the tradeoff represented in the f function within $\mathbb{P}_{Op}$.



(a) $\mathbb{R}_{Op}(\Theta)$ for the *b*Orgs and local baseline



(b) $\mathbb{C}_{Op}(\Theta)$ for the *b*Orgs and local baseline



(c) Example imparted metareasoning regimes for *b*Orgs and local baseline as a function of tradeoff parameterization

Figure 4

6 Related Work

Our work in this paper resides in the intersection of three fields of study: multiagent metareasoning, organizational modeling, and multiagent sequential decision making. Multiagent metareasoning (see Cox and Raja (2011) for a contemporary snapshot of the field) has largely treated the problem as a decentralized coordination problem, where agents model each others' reasoning processes and pass pertinent information among themselves so as to decide how best to coordinate the use of their reasoning resources. In contrast, the work we present here centralizes the problem in the ODP, amortizing the costs associated with centralization by constructing long-term metareasoning regimes about which parts of the joint problem are worthwhile for each agent to reason about given its ongoing organizational role. Thus, for specific problem instances, our algorithm's resulting solution might be suboptimal both in terms of circumscribing agents' reasoning and coordinating the timing of that reasoning, but for long-term systems, the amortized computational cost for our ODP to design an organization could be much less than the agents' costs to determine how to balance their reasoning and behaviors for each individual problem instance.

Organizational modeling research (see Dignum and Padget (2012) for a comprehensive overview) has typically focused on how to marshal agents to work together to achieve collective goals none would have achieved alone, by defining the roles, norms, interaction protocols, etc. that agents should follow. The organizational constructs thus might focus agents on considering particular tasks and interactions, and hence might simplify their reasoning. In this context, the work that we describe here, while so far lacking in the richness of modeling constructs considered in much organizational modeling research, provides a basis for raising this otherwise overlooked impact of an organizational design on agent reasoning to explicit consideration.

Given the general intractability of optimally solving decentralized decision problems, multiagent sequential decision making research has investigated a variety of algorithmic techniques for approximating, simplifying, and decoupling agents' reasoning (Hansen and Zilberstein 2001a; Witwicki and Durfee 2010; Velagapudi et al. 2011; Durfee and Zilberstein 2012; Oliehoek et al. 2013; Zhang and Lesser 2013). Rather than directly contributing to this body of techniques, our work instead emphasizes a strategy for analyzing patterns of joint behavior to selectively modify the problems agents solve. This idea has been used before to bias agents to separately find solutions that have joint benefit, through, for example, reward shaping (Agogino and Tumer 2005), transition shaping (Witwicki and Durfee 2010), organizational influence (Sleight and Durfee 2013), and hand-encoded domain knowledge (Oliehoek, Whiteson, and Spaan 2013) but that prior work did not explicitly factor quantitative impacts on agent reasoning when designing modifications to agents' local models. Recent work on optimal reward functions (Bratman et al. 2012) is a single-agent exception in that it does shape reward functions to fit an agent's cognitive limitations, but that work optimizes behavior given fixed limitations, rather than balancing behavioral benefits against reasoning costs as in our approach.

7 Conclusion

In this paper, we leveraged the Dec-MDP formalism to characterize a quantitative organizational design problem that explicitly considers both the performance of the agents' behaviors and their computational costs. We presented techniques for efficiently, approximately solving this problem through the use of incremental search, and showed how an ODP can compute the expected incremental impact of an individual influence. Our empirical evaluation confirmed that our algorithm creates organizational designs that impart a target metareasoning regime upon the agents. In the future, we plan to expand our algorithm to other organizational influence forms such as shaping the agents' state representations, transition functions, and reward functions.

8 Acknowledgments

We thank the anonymous reviewers for their thoughtful comments, and our collaborators at the University of Massachusetts for their feedback in the preliminary stages of this research. This work was supported in part by NSF grant IIS-0964512.

References

Agogino, A. K., and Tumer, K. 2005. Multi-agent reward analysis for learning in noisy domains. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, 81–88.

Alexander, G.; Raja, A.; Durfee, E. H.; and Musliner, D. J. 2007. Design paradigms for meta-control in multi-agent systems. In *Proceedings of AAMAS 2007 Workshop on Metareasoning in Agent-based Systems*, 92–103.

Becker, R.; Zilberstein, S.; Lesser, V.; and Goldman, C. V. 2004. Solving transition independent decentralized Markov decision processes. *Journal of Artificial Intelligence Research* 22(1):423–455.

Bratman, J.; Singh, S.; Sorg, J.; and Lewis, R. 2012. Strong mitigation: Nesting search for good policies within search for good reward. In *Proceedings of the Eleventh International Conference on Autonomous Agents and Multiagent Systems*, 407–414.

Cox, M. T., and Raja, A. 2011. *Metareasoning: Thinking About Thinking*. MIT Press.

Dignum, V., and Padget, J. 2012. Multiagent organizations. In Weiss, G., ed., *Multiagent Systems*. MIT Press.

Dignum, V.; Vázquez-Salceda, J.; and Dignum, F. 2005. Omni: Introducing social structure, norms and ontologies into agent organizations. In *Programming Multi-Agent Systems*. Springer. 181–198.

Durfee, E. H., and Zilberstein, S. 2012. Multiagent planning, control, and execution. In Weiss, G., ed., *Multiagent Systems*. MIT Press.

Hansen, E. A., and Zilberstein, S. 2001a. LAO*: A heuristic search algorithm that finds solutions with loops. *Artificial Intelligence* 129(1):35–62.

Hansen, E. A., and Zilberstein, S. 2001b. Monitoring and control of anytime algorithms: A dynamic programming approach. *Artificial Intelligence* 126(1):139–157.

Horling, B., and Lesser, V. 2008. Using quantitative models to search for appropriate organizational designs. *Autonomous Agents and Multiagent Systems* 16(2):95–149.

IBM. 2012. IBM ILOG CPLEX. See http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/.

Kallenberg, L. C. M. 1983. *Linear Programming and Finite Markovian Control*. Mathematical Centre Tracts.

Littman, M.; Dean, T.; and Kaelbling, L. 1995. On the complexity of solving Markov decision problems. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 394–402.

Oliehoek, F. A.; Spaan, M. T. J.; Amato, C.; and Whiteson, S. 2013. Incremental clustering and expansion for faster optimal planning in decentralized POMDPs. *Journal of Artificial Intelligence Research* 46:449–509.

Oliehoek, F. A.; Whiteson, S.; and Spaan, M. T. J. 2013. Approximate solutions for factored Dec-POMDPs with many agents. In *Proceedings of the Twelfth International Conference on Autonomous Agents and Multiagent Systems*, 563–570.

Pacheco, O., and Carmo, J. 2003. A role based model for the normative specification of organized collective agency and agents interaction. *Autonomous Agents and Multi-Agent Systems* 6(2):145–184.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc.

Raja, A., and Lesser, V. 2007. A framework for meta-level control in multi-agent systems. *Autonomous Agents and Multi-Agent Systems* 15(2):147–196.

Shoham, Y., and Tennenholtz, M. 1995. On social laws for artificial agent societies: Off-line design. *Artificial Intelligence* 73(1-2):231–252.

Sleight, J., and Durfee, E. H. 2013. Organizational design principles and techniques for decision-theoretic agents. In *Proceedings of the Twelfth International Conference on Autonomous Agents and Multiagent Systems*, 463–470.

Velagapudi, P.; Varakantham, P.; Sycara, K.; and Scerri, P. 2011. Distributed model shaping for scaling to decentralized POMDPs with hundreds of agents. In *Proceedings of the Tenth International Conference on Autonomous Agents and Multiagent Systems*, 955–962.

Witwicki, S. J., and Durfee, E. H. 2010. Influence-based policy abstraction for weakly-coupled Dec-POMDPs. In *Proceedings of the Twentieth International Conference on Automated Planning and Scheduling*, 185–192.

Zhang, C., and Lesser, V. 2013. Coordinating multi-agent reinforcement learning with limited communication. In *Proceedings of the Twelfth International Conference on Autonomous Agents and Multiagent Systems*, 1101–1108.