Efficient Metadeliberation Auctions

Ruggiero Cavallo and **David C. Parkes**
School of Engineering and Applied Sciences
Harvard University
{cavallo, parkes}@eecs.harvard.edu

Abstract

Imagine a resource allocation scenario in which the interested parties can, at a cost, individually research ways of using the resource to be allocated, potentially increasing the value they would achieve from obtaining it. Each agent has a private model of its research process and obtains a private realization of its improvement in value, if any. From a social perspective it is optimal to coordinate research in a way that strikes the right tradeoff between value and cost, ultimately allocating the resource to one party– thus this is a problem of *multi-agent metadeliberation*. We provide a reduction of computing the optimal deliberation-allocation policy to computing Gittins indices in multi-armed bandit worlds, and apply a modification of the *dynamic-VCG* mechanism to yield truthful participation in an *ex post* equilibrium. Our mechanism achieves equilibrium implementation of the optimal policy even when agents have the capacity to deliberate about other agents' valuations, and thus addresses the problem of *strategic deliberation*.

Introduction

Imagine a group of firms competing for the allocation of a new technology. Each firm may initially have some estimate of how valuable the technology is to its business, and be able to learn new ways of using the technology for greater profit through research. If such research were costless and instantaneous, the socially optimal plan would have all firms research the technology in all ways possible, at which point it would be allocated to the firm with highest value. But in reality performing such research will come at a cost. To maximize expected social welfare an optimal tradeoff should be struck between value and cost, with firms following a coordinated research policy. In addition to gathering information from the outside world, participants may improve their values for the resource by performing some costly computation, for instance finding better business plans involving the resource. We adopt the general term *deliberation* for any such value-improving process, and we consider the social planner's *metadeliberation* problem—deciding when and how to perform deliberation (including when to stop and allocate the resource.)

The main contributions of this paper lie, first, in describing a method of reducing such deliberation-allocation problems to the *multi-armed bandit* problem (see, e.g., (Gittins 1989)), thus providing a computationally efficient way of determining optimal policies. This is non-trivial because the local problem of each agent (or firm) includes two actions in each state—deliberation and allocation—and is thus not modeled as a simple Markov chain. Our setting is that of a multi-agent system with private information and self-interest. The second contribution is in applying recent work in *dynamic mechanism design* (Bergemann & Valimaki 2006) to achieve equilibrium implementation in the face of selfish, strategic parties. Our solution provides a *metadeliberation auction*, in which agents will choose to reveal private information about their deliberation processes and also to voluntarily perform deliberation as and when specified by the optimal solution.

In an extension, we allow agents to have deliberation processes for the value of other agents for the resource. This borrows from the earlier model of Larson and Sandholm (2005), in which agents have costly deliberation processes and can perform "*strategic deliberation*" about the value of other agents. But whereas they exclude solutions in which the mechanism is actively involved in coordinating the deliberation of agents, we allow for this and obtain positive results where they have impossibility results. In particular, when the optimal policy calls for one agent to perform research on behalf of another, we can achieve this. In our mechanism an agent is paid for increasing (via its deliberation process) the value of the item to another agent, and thus enjoys the beneficial results of the deliberation it performs.

Related work. On the policy computation side, the most important result for our purposes is that of Gittins and Jones (1974), who showed that the multi-armed bandit problem has a solution with complexity that grows linearly in the number of arms (we describe the result in some detail later on). Glazebrook (1979) extended this result to "stoppable" bandits, where execution of the system can be halted for a final reward. Our multi-agent deliberation-allocation problem falls within his framework and our reduction to the bandits problem is a special case of his reduction. This noted, we provide a new proof that elucidates the reduction and leverages the special structure in our environment.

Cavallo et al. (2006) proposed a dynamic mechanism suitable to this kind of environment, in which each agent's local problem is modeled as a Markov decision process (MDP), with the MDPs coupled by joint actions. Bergemann and Välimaki (2006) independently proposed the *dynamic-VCG* mechanism, providing stronger participation properties than the Cavallo et al. solution.[1]

Also strongly related is Weitzman's (1979) foundational result on optimal search among a set of alternatives, where the exact value of an alternative is revealed for a cost. Like Gittins, Weitzman develops an index policy for his search problem; however, his problem is different in that it is stoppable (like ours and that of Glazebrook) and can be applied to an undiscounted setting. Bergemann and Välimäki (2002) look at the problem of information acquisition by bidders in a single-item auction, and show that when such acquisition is one-shot and simultaneous among the group, the Vickrey auction provides the right *ex ante* incentives. Larson (2006) and Cremer et al. (2007) use Weitzman's result to form an optimal-search auction model with sequential information acquisition, but also assume that a buyer's acquisition process is instantaneous (not multi time-stepped, with incremental information). Parkes (2005) studied the role of auction design given participants that have costly or limited value refinement capabilities, especially the tradeoff between sealed bid and iterative designs, but does not provide an optimal design.

The setting

Members of a set I of n agents ($\{1, 2, \ldots, n\}$) compete for allocation of a resource. Each agent $i \in I$ has an initial value for the resource, and can refine its value repeatedly via costly "deliberation". To keep things simple we will initially assume that each agent has only one such deliberation process, and moreover that no agent has a deliberation process about the value of any other agent.

Each agent i has an MDP model $M_i = (S_i, A_i, \tau_i, v_i, c_i)$ for how its valuation for the resource will change subject to deliberation. S_i is i's local state space. The action space $A_i = \{\alpha_i, \beta_i\}$, where α_i allocates the resource to i and β_i is deliberation by i. We use $v_i(s_i) \geq 0$ to denote the value an agent would obtain from being allocated the resource (performing no additional deliberation) in state $s_i \in S_i$. $c_i \geq 0$ is the cost each agent i incurs every time it performs its deliberation action. For simplicity we assume c_i is constant, though our results hold as long as c_i is a bounded function of i's state. States evolve according to a (possibly nondeterministic) transition function, $\tau_i(s_i, a_i) \in S_i$, defined so that $\tau_i(s_i, \alpha_i) = \phi_i$, where $\phi_i \in S_i$ is a special *absorbing* state from which no additional actions are available. Agents have a common discount factor γ, where $0 \leq \gamma < 1$.

A set of further assumptions placed on this framework defines a *domain*. In our setting, researching new uses may yield a greater value for the resource, but agents won't forget previously known uses, so the following is natural:

Assumption 1 (Uncertainly improvable values). *Agent valuations never decrease:* $v_i(\tau_i(s_i, \beta_i)) \geq v_i(s_i)$, $\forall s_i \in S_i$.

Consider the agent MDP represented in Figure 1. If the agent deliberates once, with probability 0.33 its valuation for the resource (i.e., the value it would obtain if allocated) will increase from 0 to 3, and with probability 0.67 it will increase only to 1. If a second deliberation action is taken and the current value is 1, with equal probability the valuation will stay the same or increase to 4; if the current value is 3, it will increase to 4 with certainty.

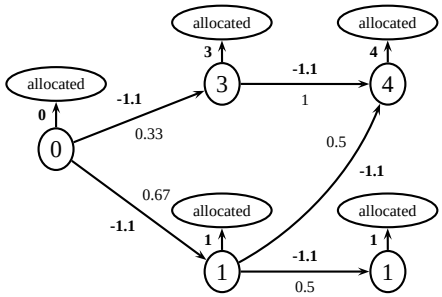


Figure 1: Example of an agent's MDP model of how its valuation for the resource would change upon deliberation, labeled with transition probabilities and instantaneous rewards (in bold). The agent's cost of deliberation is 1.1.

We make the following additional assumptions:

Assumption 2. *Agent deliberation processes are independent of each other.*

Assumption 3. *Agents cannot deliberate prior to the beginning of the mechanism.*

Assumption 4. *Only one action can be taken per time-step (i.e., multiple agents cannot deliberate concurrently).*

Assumption 2 is already implicit in our setup, with agent transitions and rewards functions of only *local* states, and actions for one agent causing transitions only in its own MDP. Assumption 3 can be motivated by considering that the resource is "revealed" only at the beginning of the mechanism. Finally, Assumption 4 is without loss of generality when the discount factor is high enough because it would be socially optimal to deliberate sequentially in that case anyway.

Combining the agent problems, we have a multi-agent MDP (see Boutilier (1996)) $M = (S, A, \tau, v, c)$ in which $S = S_1 \times \ldots \times S_n \times \{0, 1\}$ and $A = A_1 \cup A_2 \cup \ldots \cup A_n$, with $\tau(s, a) = (s_1, \ldots, \tau_i(s_i, a), \ldots, s_n, 0)$ if $a = \beta_i$ and $\tau(s, a) = (s_1, \ldots, \phi_i, \ldots, s_n, 1)$ if $a = \alpha_i$, i.e. the final bit in the state (denoted $\Lambda(s) \in \{0, 1\}$) indicates whether or not the resource has been allocated. Notation v and c denote a valuation profile $(v_1, \ldots, v_n)$ and cost profile $(c_1, \ldots, c_n)$ respectively.[2] Given this, we define reward function $r(s, a)$

[1] More recently, Cavallo et al. (2007) have generalized these methods, providing mechanisms that allow for agent arrivals and departures, and also periods of inaccessibility. Also related is Ieong et al.(2007).

[2] We assume that each agent has a correct model for its local

for the multi-agent MDP as $\sum_{i \in I} r_i(s,a)$, with, $\forall i \in I, s \in S, a \in A$:

$$r_i(s,a) = \begin{cases} 0 & \text{if } \Lambda(s) = 1, \text{ or } a \notin \{\alpha_i, \beta_i\} \\ v_i(s_i) & \text{if } \Lambda(s) = 0 \text{ and } a = \alpha_i \\ -c_i & \text{if } \Lambda(s) = 0 \text{ and } a = \beta_i, \end{cases}$$

This captures the essential aspect of the problem: the process "stops" once the resource has been allocated, and upon allocation the agent that receives the item obtains the value associated with its current state.

Consider decision policy π, where $\pi(s) \in A$ is the action prescribed in state s. We write $V_i^{\pi}(s^0) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[r_i(s^t, \pi(s^t))], \forall s^0 \in S$, where $s^t = \tau(s^{t-1}, \pi(s^{t-1}))$ for $t > 0$. We write $V^{\pi}(s) = \sum_{i \in I} V_i^{\pi}(s), \forall s \in S$. Let π^* denote the socially optimal policy, i.e., $\pi^* = \arg\max_{\pi \in \Pi} V^{\pi}(s), \forall s \in S$, where Π is the space of all action policies. We will at times consider a policy $\pi_i^* : S_i \rightarrow A$ that is optimal for agent i, i.e., $\pi_i^* = \arg\max_{\pi \in \Pi} V_i^{\pi}(s)$, $\forall s \in S$. We use $V^*(s)$ as shorthand for $V^{\pi^*}(s)$, and $V_i^*(s_i)$ for $V_i^{\pi_i^*}(s)$. Letting Π_{-i} denote the policies that never choose deliberation or allocation for i (as though i were not present in the system), we write $\pi_{-i}^* = \arg\max_{\pi \in \Pi_{-i}} V_{-i}^{\pi}(s)$, and $V_{-i}^*(s_{-i})$ as shorthand for $V_{-i}^{\pi_{-i}^*}(s)$. We also define, $\forall s \in S, a \in A$:

$$Q(s,a) = \sum_{i \in I} r_i(s_i, a) + \gamma \mathbb{E}[V^*(\tau(s,a))],$$

$$Q_i(s_i, a) = r_i(s_i, a) + \gamma \mathbb{E}[V_i^*(\tau(s_i, a))], \text{ and}$$

$$Q_{-i}(s_{-i}, a) = \sum_{j \in I \setminus \{i\}} r_j(s_j, a) + \gamma \mathbb{E}[V_{-i}^*(\tau(s_{-i}, a))]$$

This formulation is quite general and allows, for example, for the local states to represent "information states" in the sense of models of optimal Bayesian learning (Bellman & Kalaba 1959), as well as performance profile trees of the form proposed by Larson and Sandholm for normative metadeliberation (2001) (with the added restriction that values cannot decrease).

We consider procedures in which agents report state and MDP information to a "*center*," such as an auctioneer. The center executes a deliberation-allocation policy, in each period either suggesting to some agent that it take a deliberation action or allocating the resource (and ending the process). Agents are *self-interested* and may subvert the process by misreporting information or by not following a deliberation action suggested by the center. Before presenting our solutions, we give a brief background on the methods we leverage for efficiently computing optimal policies and managing the self-interest of agents.

Background: Policies for multi-armed bandits

In multi-armed bandit (MAB) problems, there is a set of n reward-generating Markov processes, $\{1, \ldots, n\}$, and exactly one process can be activated every time-step. The reward that a process i generates if activated at time t is a function only of its state s_i^t at t (and not of any other process's state). If i is chosen at t, a reward $r_i(s_i^t)$ is obtained and successor state s_i^{t+1} is reached (perhaps non-deterministically) according to s_i^t; for all $j \neq i$, $s_j^{t+1} = s_j^t$ and no reward is generated at t.

Theorem 1. (Gittins & Jones 1974; Gittins 1989) *Given Markov chains $\{1, \ldots, n\}$, joint state space $S = S_1 \times \ldots \times S_n$, discount factor $0 \leq \gamma < 1$, and an infinite time-horizon, there exists a function $\nu : S_1 \cup \ldots \cup S_n \rightarrow \mathcal{R}$ such that the optimal policy $\pi^*(s) = \arg\max_i \nu(s_i), \forall s \in S$.*

So the complexity of computing an optimal policy is linear in the number of processes. In contrast, general multi-agent MDP problems scale exponentially in the number of agents because of the size of the joint state space. Such a function ν has come to be called the "Gittins index." Several methods for computing Gittins indices are known.[3]

But our problem is not quite a bandits problem. If our agents are considered the arms of the MAB problem, each arm has *two* local actions—allocate and deliberate—and is not a Markov chain. There is also special structure to our problem: if an allocation action α_i is taken then the whole system stops. Glazebrook (1979) considered a similar setting, in which the local MDP for each arm could be reduced to a Markov chain by pre-solving for the optimal local policy, supposing that the arm was activated in every time-step. This approach also applies here: his "condition (b)" is our uncertainly improvable values (UIV) condition, which will allow us to prune away one action from every state of an agent's local MDP, yielding Markov chains. We thus reduce the problem to a multi-armed bandit, which is then solvable via Gittins indices. We offer an independent proof of Glazebrook's result, exposing additional structure of the problem when this UIV property holds; Glazebrook's proof is for a more general condition, later shown to be implied by his condition (b).

Background: Dynamic mechanism design

Mechanism design addresses the problems posed by self-interest by adjusting agent payoffs, via *transfer payments*, such that each agent's utility is maximized exactly when social utility is maximized. We adopt the framework of *dynamic mechanism design*, which is applicable to sequential decision-making settings in which agents obtain new information over time (see Parkes (2007) for a recent survey). The following steps occur in each period: the agents report claims about private information; the center takes or proposes actions; the agents take actions; the center makes payments to the agents. This process is repeated over and over until an allocation is finally made.

The *dynamic-VCG mechanism* (Bergemann & Valimaki 2006) provides the basis for our solution and is defined for "private" multi-agent MDP models such as ours in which local agent rewards and transitions are independent of the

deliberation process; from this the multi-agent MDP is also correct.

[3]For instance, Katehakis and Veinott (1987) provide a way of defining a "restart-in-s_i" MDP for each process i, in any state s_i, the value of which is equivalent to the Gittins index for the process in that state.

states of other agents when one conditions on actions by the center. Dynamic-VCG is defined as follows: at every time-step t, the socially optimal decision according to reported joint state s^t and reported MDP models is chosen, and each agent i is payed: $Q^*_{-i}(s^t_{-i}, \pi^*(s^t)) - V^*_{-i}(s^t_{-i})$. Intuitively, at each time-step each agent must pay (reversing the signs in this equation) the center a quantity equal to the extent to which its current report inhibits other agents from obtaining value in the present and in the future.

A *within-period ex post Nash equilibrium* is one in which, at every time-step, for any joint state, every agent maximizes utility by playing its equilibrium strategy when others do (now and in the future). A mechanism is *incentive-compatible* in this equilibrium if agents are best off reporting private information truthfully and acting according to the center's prescriptions when others do (whatever their private information), and *individual rational (IR)* if expected payoff is non-negative to an agent playing the equilibrium strategy when others do.

Theorem 2. (Bergemann & Valimaki 2006) *The dynamic-VCG mechanism for private multi-agent MDPs is optimal, incentive compatible, and IR in a within-period ex post Nash equilibrium, and never runs a deficit.*[4]

Results: Efficient Computation

Our computational approach is to reduce each agent's local MDP to a local Markov Chain (MC) by pruning one of the allocate/deliberate actions in each state, which will then allow for an index policy that is optimal also for the unpruned MDPs. Recalling the MDP model for a single agent depicted in Figure 1; Figure 2 portrays the same MDP after the pruning away of actions that would not be optimal *in a world in which the agent existed alone*.

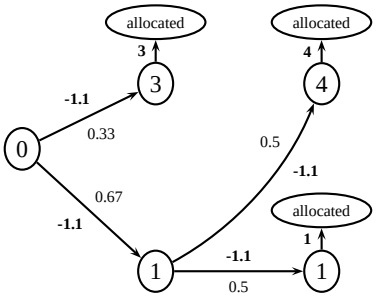


Figure 2: MDP world model from Figure 1 for a single agent i, after pruning of actions that would not be optimal in a world with no other agents. $\gamma = 0.95$, $c_i = 1.1$.

Definition 1 (Efficiently MC-prunable). *A domain is efficiently MC-prunable if and only if, for any agent i, for any agent MDP models, any action that would not be optimal in a world with no agents other than i is not socially-optimal in the multi-agent MDP problem, i.e.,*

$$\forall i \in I, \forall a_i \in \{\alpha_i, \beta_i\}, \forall s \in S, \;\; a_i \notin \pi^*_i(s_i) \Rightarrow a_i \notin \pi^*(s) \quad (1)$$

We will, for our domain, establish this property and subsequently the validity of the following procedure:

- Convert each agent's MDP model into a Markov chain by determining the policy that would be optimal if no other agents were present.
- Perform the deliberation-allocation process, computing an index for each agent MC at every time-period, always activating an MC with highest index.

The following lemma shows that to test for efficient MC-prunability in our domain, we can restrict our analysis to the pruning of deliberation actions.

Lemma 1. *A domain is efficiently MC-prunable if and only if*

$$\forall i \in I, \; \forall s \in S, \;\; \beta_i \in \pi^*(s) \Rightarrow \beta_i \in \pi^*_i(s_i) \quad (2)$$

Proof. Considering the contrapositive of (1), efficient MC-prunability requires that (2) and the following hold:

$$\forall i \in I, \; \forall s \in S, \;\; \alpha_i \in \pi^*(s) \Rightarrow \alpha_i \in \pi^*_i(s_i) \quad (3)$$

It turns out that (3) holds for any domain. Observe that $Q(s,a) \geq Q_i(s_i,a), \forall a \in A$, as π^* is optimized over policy space Π, and $\pi^*_i \in \Pi$. Assume that $\alpha_i \in \pi^*(s)$ and, for contradiction, that $\alpha_i \notin \pi^*_i(s)$, i.e., that $Q(s,\alpha_i) \geq Q(s,a), \forall a \in A$, and $Q_i(s,\beta_i) > Q_i(s,\alpha_i)$. We have:

$$Q(s,\alpha_i) \geq Q(s,\beta_i) \geq Q_i(s_i,\beta_i) > Q_i(s_i,\alpha_i) = Q(s,\alpha_i),$$

a contradiction. □

In the full version of this paper, we use the above characterization to prove the following lemma.

Lemma 2. *All uncertainly improvable values domains are efficiently MC-prunable.*

This enables a "without loss" reduction from local MDPs to local MCs. The remaining challenge is that the Gittins index policy is only optimal for problems with an infinite time-horizon. This issue can be handled when $\gamma < 1$ by replacing the one-time reward of $v_i(s_i)$ in a state s_i in which agent i is allocated the item with a reward of $(1-\gamma)v_i(s_i)$ received per period in perpetuity. It is then a simple matter to show that the optimal MAB policy will always continue to activate agent i's MC after it first does so when i is in an "allocation state". Thus the resulting policy is valid for the original problem with absorbing states. Returning to our example, Figure 3 displays the infinite horizon, pruned MC for the problem earlier depicted in Figure 2.

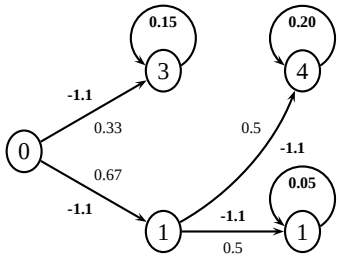


Figure 3: Agent-optimal Markov chain from Figure 2 after expansion to an infinite-horizon.

Theorem 3. *Given Assumptions 1–4, the deliberation-allocation policy defined by activating, at every time-step t, the pruned, locally-optimal Markov chain of an agent with the highest Gittins index is optimal.*

[4] See Cavallo et al.(2007) for a simple proof.

Results: Handling Selfish Agents

We now combine the index-policy solution to the multi-agent metadeliberation problem with the dynamic-VCG mechanism to obtain our *metadeliberation auction*, in which the center chooses actions based on private information that agents report. Note that, in the case of a deliberation action, "chooses" means "suggests to the agents"; for an allocation action, the center simply executes it.

Mechanism 1 (Metadeliberation auction). .

- *Each agent i computes its locally optimal, infinite-horizon Markov chain M_i^*, and reports to the center claims $\hat{M}_i^*$ and $\hat{s}_i^0$ about M_i^* and initial local state s_i^0.*
- *At every time-step t (with agents in true state s^t), while the resource has not yet been allocated:*
  1. *The agent i activated in the previous time-step reports a claim $\hat{s}_i^t$ about its current state.*[5]
  2. *The center chooses the action specified by activation of an agent i^* with highest Gittins index.*
  3. *Agent i^* pays the center:*

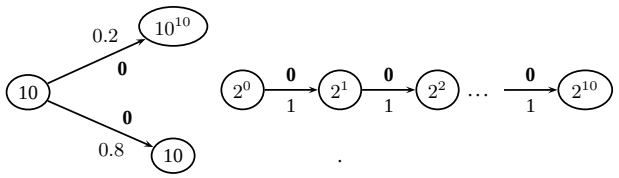
$$\begin{array}{ll} (1-\gamma)\, V^*_{-i^*}(\hat{s}^t_{-i^*}) & \text{if deliberation was performed,} \\ V^*_{-i^*}(\hat{s}^t_{-i^*}) & \text{if the item was allocated} \end{array}$$

Theorem 4. *Given Assumptions 1–4, Mechanism 1 is optimal, incentive compatible, and IR in a within-period ex post Nash equilibrium, and never runs a deficit.*

Proof. The result follows from Theorems 1, 2, and 3. The dynamic-VCG mechanism requires that each agent i pay the center an amount equal to the negative externality its presence imposes on the other agents at t, i.e., $V^*_{-i}(\hat{s}^t_{-i}) - Q^*_{-i}(\hat{s}^t_{-i}, \pi^*(\hat{s}^t))$. In our setting, for the agent who deliberates at t this is equal to the cost to the other agents of having to wait one time-step to implement the policy that would be optimal for them, i.e., $(1-\gamma)\, V^*_{-i^*}(\hat{s}^t_{-i^*})$; for all other agents it is 0. When the item is allocated to an agent, that agent imposes an externality equal to the total value agents could get from the current state forward if he were not present. □

This provides the result we want: each agent will first prune away its suboptimal local actions, and then truthfully report its (pruned) MC to the center. From that point forward, the center will suggest deliberation actions according to the optimal deliberation-allocation policy, collecting a payment from the agent that deliberates. Agents will choose to follow these suggestions and truthfully report new local states, and the center will eventually allocate the resource. At that point the agent will consume the resource with no further deliberation, by optimality of the deliberation-allocation policy.

Example 1 Consider the execution of Mechanism 1 on the example in Figure 4 (for simplicity we've switched to a more concise representation, omitting allocation nodes). The optimal policy has agent 1 deliberate first; if his value increases to 10^{10} he is then allocated the resource. Otherwise the optimal policy has agent 2 deliberate for 10 time-steps and then allocates to him. Under Mechanism 1, in the first time-step agent 1 must pay the "immediate externality" imposed on agent 2 assuming the policy optimal for agent 2 would be executed in all following periods, i.e., his cost of waiting one period, or $(1-0.9)\cdot 0.9^{10}\cdot 2^{10}$. If agent 1's deliberation yields the high value (10^{10}) he must then pay $0.9^{10}\cdot 2^{10}$.



(a) Agent 1 (b) Agent 2

Figure 4: Agent 1 has initial value 10. With small probability his value will increase to 10^{10} if he deliberates once. Agent 2's value is $\min(2^x, 2^{10})$, where x is the number of times he has deliberated. $c_1 = c_2 = 0$ and $\gamma = 0.9$.

If agent 1's deliberation does not yield the improvement, then in every period that follows prior to allocation (with agent 2 deliberating) agent 2 must pay $(1-0.9)\cdot 10 = 1$. In the final allocation step agent 2 pays 10. Bear in mind that agent 2 *discounts* value (rewards and costs) in the future by a factor of 0.9. We can compute agent 2's expected utility (from the first time he is asked to deliberate) for being truthful, and compare it to his expected utility if he misreports his MC such that he is asked to instead deliberate for only $k < 10$ time-steps, and then finishes his deliberation once he receives the resource. For any k, if agent 2 deliberates k times, the total discounted payments he makes will equal:

$$\begin{aligned} &(1-\gamma)10 + \gamma(1-\gamma)10 + \ldots + \gamma^{k-1}(1-\gamma)10 + \gamma^k 10 \\ &= 10 - \gamma 10 + \gamma 10 - \ldots - \gamma^{k-1}10 + \gamma^{k-1}10 - \gamma^k 10 + \gamma^k 10 \\ &= 10 \end{aligned}$$

So his *discounted payments* are the same regardless of how many times he deliberates. Then since it is optimal for agent 2 to deliberate 10 times, whether he does so inside or outside the context of the mechanism, his total discounted utility will always equal $\gamma^{10}2^{10} - 10$.

Extensions: Multiple deliberation processes

So far, in order to simplify the analysis we've assumed that each agent has only one way of deliberating. However, our results also apply when agents have multiple independent deliberation methods. For instance, imagine an agent that has three different research programs it could pursue (potentially with distinct associated costs per time-step)—the agent merely has to report all three models to the center, who will consider all three in determining the optimal policy. It is important, though, that all deliberation processes are *independent* (deliberation in one process cannot change the state of another process); otherwise, there will be no reduction to the multi-armed bandit problem. Given this independence, a generalization of Theorem 4 immediately follows.

[5] Technically, at every time-step each agent must have a chance to report a new claim about its Markov chain model and current state, but this presentation is consistent with the equilibrium behavior (the same applies to Mechanism 2).

Strategic Deliberation

Consider now a setting in which an agent may have one or more deliberation processes that pertain to the value of *other* agents for the resource. This models the setting of strategic deliberation introduced by Larson and Sandholm (2001).[6] We retain the ability to implement optimal deliberation-allocation policies in this context. Note that the optimal policy might specify "cross-agent" deliberation, with the results of i's research being shared with j (in particular, when i has a better deliberation process than j).

The dynamic-VCG scheme *will not* work here. A subtle condition usually required for the good incentive and IR properties of dynamic-VCG is that the optimal policy for agents other than i does not take any actions that involve agent i. Formally, the necessary condition is that $\max_{\pi \in \Pi} V^{\pi}_{-i}(s) = \max_{\pi \in \Pi_{-i}} V^{\pi}_{-i}(s)$ (see Cavallo et al. (2007) for a discussion). This condition is not met when the optimal policy has one agent deliberate about another's value. The intuition behind the extension of dynamic-VCG that we present is that the payments make the expected equilibrium payoff to agent i forward from any state equal to the payoff i would receive in the dynamic-VCG mechanism *if its deliberation processes about other agents were actually about itself.* The equilibrium properties then follow immediately from the analysis of Mechanism 1 in the context of agents with multiple independent deliberation processes.

Let p_{ij} denote a deliberation process possessed by agent i pertaining to the value agent j would achieve from the resource; we let $c_{p_{ij}}$ denote the cost (to i) of deliberating on process p_{ij}. For any process p_{ij}, any state $s_{p_{ij}}$ consists of two things: some *information* $I(s_{p_{ij}})$ (e.g., the observations of the world acquired from research, or the plan resulting from some computation), and a valuation $v(s_{p_{ij}})$ for j receiving the item given the information content. Let $v_j(I(s_{p_{ij}}))$ denote the *actual* value received by j for the information associated with the same state. Allowing for misreports, $v(\hat{s}_{p_{ij}})$ denotes the value that should be achieved by j according to i's state report, $I(\hat{s}_{p_{ij}})$ denotes the information content associated with that state report, and $\hat{v}_j(I(\hat{s}_{p_{ij}}))$ is a claim made by j about the actual value it achieved. In Mechanism 2 the center computes payments by reasoning about the social value that could be achieved under a policy that is optimal with all agents present, but in which an agent i *cannot take any actions*. We denote this quantity, which is independent of i's state, as $V^{*-i}(s_{-i})$, for all $s \in S$.

Theorem 5. *Given Assumptions 1–4, Mechanism 2 is optimal, incentive compatible, and IR in a within-period ex post Nash equilibrium, and does not run a deficit when agents are truthful.*

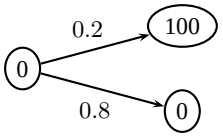
Proof Sketch. The incentive and IR properties of the mechanism follow from those of Mechanism 1, combined with the following observation: for any process p_{ij} with $i \neq j$, the payment scheme yields a scenario which is, payoff-wise, identical to one in which p_{ij} is a deliberation process pertaining to i's value. If p_{ij} is selected for deliberation then i already pays the cost. If p_{ij} is selected for allocation then i will be paid an amount equal to the actual value yielded from the process (assuming agent j is honest), and j will obtain value 0 (assuming i is honest), since $v(s^t_{p_{ij}}) = v_j(I(s^t_{p_{ij}}))$ by our assumption that beliefs are correct.[7] The mechanism never runs a deficit in equilibrium. Prior to the final allocation step there are no payments that flow to the agents. Then in that final allocation step payments made to the center are $v(s^t_{p_{ij}}) + V^{*-i}(\hat{s}^t_{-i}) - v_j(I(s^t_{p_{ij}}))$. Given truthful reporting (which, as shown above, is achieved in an ex post equilibrium), this quantity equals $V^{*-i}(\hat{s}^t_{-i})$, which is ≥ 0. □

Mechanism 2 (with cross-agent deliberation). .

- *Each agent i computes the locally optimal, infinite-horizon Markov chain for every deliberation process it possesses, and reports claims about each MC and initial local state to the center.*
- *At every time-step t (with agents in true state s^t), while the resource has not yet been allocated:*
  1. *For process $p_{\bar{i},\bar{j}}$ activated in the previous time-step, agent i reports a claim $\hat{s}^t_{p_{\bar{i},\bar{j}}}$ about $p_{\bar{i},\bar{j}}$'s current state.*
  2. *The center chooses the action specified by activation of a process p_{ij} with highest Gittins index.*
  3. *If deliberation was performed, agent i pays the center $(1-\gamma)\, V^{*-i}(\hat{s}^t_{-i})$.*

     If the item was allocated and $i = j$, j pays the center $V^{*-j}(\hat{s}^t_{-j})$. If $i \neq j$, the center communicates $I(\hat{s}^t_{p_{ij}})$ to agent j, j communicates $v_j(I(\hat{s}^t_{p_{ij}}))$ to the center, i pays the center $V^{*-i}(\hat{s}^t_{-i}) - v_j(I(\hat{s}^t_{p_{ij}}))$, and j pays the center $v(\hat{s}^t_{p_{ij}})$.

Example 2 Consider a 2-agent scenario in which agent 1 will obtain value 10 if allocated the resource (deliberation changes nothing), and agent 2 has one deliberation step, which yields value 100 with probability 0.2, and otherwise yields value 0. Consider that agent 1 and agent 2 each have a deliberation process with this structure *for agent 2*, but that agent 2's cost of running his is **50** and agent 1's cost is **1**. Take $\gamma = 0.9$.



(a) Agent 1 does not have an incentive to deviate from truthfulness—for instance, simply claiming agent 2 has the high 100 value without deliberating for him. Agent 1 will be payed the value that *agent 2 reports experiencing, given the information obtained from agent 1's deliberation.* So agent 1's payment is only based on agent 2's *actual* utility (assuming agent 2 is truthful). If agent 1 reported agent 2 had the

[6] But note that our independence assumption precludes results of one agent's deliberation impacting the expected results of another's, though they may concern the same agent's value.

[7] Note that if an agent i is allocated the item via an agent j's process, both agents are indifferent about their reports during the final allocation stage. Ex post IC and IR are technically maintained as there is only one "possible" true state for j, and it is known to i. There is an alternate payment scheme that avoids this indifference, but in some cases a deficit will result in equilibrium.

high value and didn't communicate corresponding information (e.g., a plan for using the resource), the value agent 2 experiences—and the value agent 1 is payed—would be 0.[8]

(b) Now consider a variant in which agent 2's cost of deliberating is **5** rather than **50**. Agent 2 knows that if he reports truthfully agent 1 will be selected first (since agent 1's deliberation process about agent 2 is superior: cost 1), and if agent 1's deliberation yields a plan worth value 100 he will obtain none of the surplus. So would he prefer to report cost 0 in order to be asked to perform the deliberation himself first? No. Mechanism 2 specifies that he would be charged *as though agent 1's deliberation processes were about agent 1*. So in the first period agent 2 would be charged $(1-\gamma)[\gamma(0.2 \cdot 100 + 0.8 \cdot 10) - 1] = 2.42$ and pay deliberation cost 5. If agent 2's deliberation yields the high value (probability 0.2) he would obtain the resource (value 100) and make payment $\gamma(0.2 \cdot 100 + 0.8 \cdot 10) - 1 = 24.2$. If it yields low value he gets 0 and pays 0. Thus agent 2's expected utility from this strategy is $-2.42 - 5 + 0.2 \cdot 0.9 \cdot (100 - 24.2) = 6.224$. But if agent 2 is truthful, he still has a chance for high payoff; recall that the two deliberation processes are *independent*, so the result of one does not impact what the result of the other will be. In particular, if agent 1 deliberates first agent 2 has expected value $\gamma 0.8(-(1-\gamma)10 + 0.2(\gamma(100-10)) + 0.8 \cdot 0 - 5) = 7.344$. (With probability 0.8 agent 1's value for agent 2 is 0, and then agent 2 is asked to deliberate and with probability 0.2 will achieve value 100, making a payment of 10.) Thus truthfulness is a superior strategy for agent 2.

To summarize: our modification of dynamic-VCG specifies cross-agent deliberation exactly when it is socially-optimal. The payments align agents' welfare with that of the whole system, so an agent's utility maximizing strategy is exactly the strategy that maximizes utility for the system, i.e., truth.

Conclusion

This paper makes two distinct contributions. First, we demonstrate that the multi-armed bandits problem is suitable for solving multi-agent metadeliberation problems, in this case by careful reduction of the original multi-agent MDP model into a multi-agent MC model. Second, we provide a novel application of the developing theory of dynamic mechanism design to coordinate deliberative processes of involved, self-interested agents, improving social welfare. This provides, to our knowledge, the first normative solution in the setting in which information acquisition by participants is incremental rather than instantaneous. We extend the intuition of the dynamic-VCG mechanism to an environment in which it cannot be directly applied because of positive externalities. Remarkably, this does not lead to a budget deficit. There are many directions for future work. Perhaps most exciting would be an extension to the undiscounted setting where agents are completely patient; no index policy is currently known for such settings. There is also another class of compelling deliberation scenarios in which deliberation yields *better estimates* of a true valuation; i.e., agents *learn* their valuations through research, rather than increase them by learning new uses for a resource. This setting is not amenable to the reduction technique applied here.

[8]Note that this is not an issue of "punishment." Rather, in equilibrium it will *never* be useful to deviate.

Acknowledgments

The authors would like to thank the three anonymous reviewers and SPC member for exceptionally detailed, constructive reviews. This work is supported in part by NSF grant IIS-0238147 and a Yahoo! faculty award.

References

Bellman, R., and Kalaba, R. 1959. A mathematical theory of adaptive control processes. *Proc. of the National Academy of Sciences of the United States of America* 45(8):1288–1290.

Bergemann, D., and Valimaki, J. 2002. Information acquisition and efficient mechanism design. *Econometrica* 70(3):1007–1033.

Bergemann, D., and Valimaki, J. 2006. Efficient dynamic auctions. Cowles Foundation Discussion Paper 1584, http://cowles.econ.yale.edu/P/cd/d15b/d1584.pdf.

Boutilier, C. 1996. Planning, learning and coordination in multiagent decision processes. In *Proceedings of the Conference on Theoretical Aspects of Rationality and Knowledge*, 195–210.

Cavallo, R.; Parkes, D. C.; and Singh, S. 2006. Optimal coordinated planning amongst self-interested agents with private state. In *Proceedings of the Twenty-second Annual Conference on Uncertainty in Artificial Intelligence (UAI'06)*.

Cavallo, R.; Parkes, D. C.; and Singh, S. 2007. Online mechanisms for persistent, periodically inaccessible self-interested agents. In *DIMACS Workshop on the Boundary between Economic Theory and Computer Science*.

Cremer, J.; Spiegel, Y.; and Zheng, C. Z. 2007. Auctions with costly information acquisition. Iowa State University Department of Economics Working Papers Series.

Gittins, J. C., and Jones, D. M. 1974. A dynamic allocation index for the sequential design of experiments. In *In Progress in Statistics*, 241–266. J. Gani et al.

Gittins, J. C. 1989. *Multi-armed Bandit Allocation Indices*. New York: Wiley.

Glazebrook, K. D. 1979. Stoppable families of alternative bandit processes. *Journal of Applied Probability* 16:843–854.

Ieong, S.; So, A. M.-C.; and Sundararajan, M. 2007. Mechanism design for stochastic optimization problems. In *3rd International Workshop on Internet and Network Economics*, 269–280.

Katehakis, M. N., and Veinott, A. F. 1987. The multi-armed bandit problem: decomposition and computation. *Mathematics of Operations Research* 22(2):262–268.

Larson, K., and Sandholm, T. 2001. Costly valuation computation in auctions. In *Eighth Conference of Theoretical Aspects of Knowledge and Rationality (TARK VIII)*.

Larson, K., and Sandholm, T. 2005. Mechanism design and deliberative agents. In *Proc. of the Fourth Int. Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2005)*.

Larson, K. 2006. Reducing costly information acquisition in auctions. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Parkes, D. C. 2005. Auction design with costly preference elicitation. *Annals of Mathematics and AI* 44:269–302.

Parkes, D. C. 2007. Online mechanisms. In Nisan, N.; Roughgarden, T.; Tardos, E.; and Vazirani, V., eds., *Algorithmic Game Theory*. CUP.

Weitzman, M. L. 1979. Optimal search for the best alternative. *Econometrica* 47:641–654.