# Scalable Algorithms for Multiagent Sequential Decision Making

**Ekhlas Sonu**
THINC Lab, Department of Computer Science
University of Georgia
Athens, GA 30602 USA
esonu@uga.edu

## Introduction

In artificial intelligence, *decision theory* deals with computing a sequence of actions (policy) that an autonomous agent must take in order to optimize its rewards (obtain its goals in the most efficient manner). In many real world situation, an autonomous agent must deal with various sources of uncertainty while computing its optimal policy. In single agent settings, such decision making problems are formalized by partially observable Markov decison processes (POMDPs) (Kaelbling, Littman, and Cassandra 2009). An agent acting alone in non-deterministic settings may face uncertainty from various sources: the underlying dynamics of the environment and its evolvement over time may be non-deterministic, the actions performed by the agent may have non-deterministic effects, and the observations received by the agent may be noisy or they may provide only partial information about the world it inhabits.

In multiagent settings, however, in addition to the aforementioned uncertainties, an agent must also consider its interactions with other agents sharing the common environment. The agents may interact through the state of the environment, the observations received, and the rewards earned – all of which could be affected by the actions of the other agents. Hence an agent must also predict the actions that the other agents are likely to take at each time step. Depending on the type of interactions between the agents, POMDPs are generalized in one of two ways: on one hand in settings where the agents share a common reward and a common prior belief (e.g. team settings) decison making is formalized by *decentralized POMDPs* (Dec-POMDPs) (Bernstein et al. 2002), on the other hand in settings where a self-interested agent must optimize its own rewards in presence of other agents that may not share common interests or common priors the decision making is formalized by *interactive POMDPs* (I-POMDPs) (Gmytrasiewicz and Doshi 2005).

My dissertation studies the decision making process for self-interested agents in multiagent settings as formalized by I-POMDPs. Particularly, I study scalable methods to tractably solve such sequential decison making problems whose complexity is doubly exponential in some variables. In recent times I-POMDPs have found a myriad of applications across several disciplines which testifies to its growing appeal. In the field of law enforcement, I-POMDPs have been used to explore strategies for countering money laundering (Meissner 2011; Ng et al. 2010). In defense, I-POMDPs have been enhanced to include trust levels for facilitating defense simulations (Seymour and Peterson 2009a; 2009b). They have been used to produce winning strategies for playing the lemonade stand game (Wunder et al. 2011), explored for use in playing Kriegspiel (Del Giudice,Gmytrasiewicz and Bryan 2009), and even discussed as a suitable framework for a robot learning tasks interactively from a human teacher (Woodward and Wood 2012; Wang 2013). I-POMDPs have been modified to include empirical models for simulating human behavioral data pertaining to strategic thought and action (Doshi et al. 2010). The growing appeal of I-POMDPs have necessitated research for exploring scalable solution algorithms for the framework.

While computing an optimal policy, in addition to the uncertainty in the evolvement of the environment and the information recieved at each step in the form of observations, an I-POMDP agent must also predict the behavior of the other agents. To do so, the subject agent must maintain a set of possible models of the other agents based on what it believes are their beliefs, desires, and intentions. The problem is further complicted as the other agents may themselves be rational agents that model all other agents (including the subject agent) in a similar manner. This complication leads to a form of recursive modeling of other agents models which in a two agent setting could be described as "what agent $i$ thinks about what agent $j$ thinks about what agent $i$ thinks, . . . , and so on". I-POMDPs limit this form of nested reasoning to a finite level in order to achieve convergence. To capture uncertainty involved in the current state of the environment, an agent maintains a joint probability distribution over the physical states and the set of models of the other agents known as its *belief*. Over time, it updates the belief using Bayesian update and computes optimal action to take according to the update belief.

Naturally, each source of uncertainty increases the complexity of decision making which are defined as its various *curses*. In my dissertation, I propose various solution techniques that address each of these curses while improving the scalability of I-POMDPs. First formally define the I-POMDP framework and describe the curses involved in solving them. Next I discuss the research I have done as a part of my dissertation and the research that I plan to finish before my defense. Finally I discuss my plans for future research, my career plans after graduation, and what I hope to gain from the doctoral consortium at ICAPS 2015.

## Background

The problem of decision making under uncertainty for a self-interested agents in multiagent settings is formalized by interactive POMDPs (I-POMDPs) (Gmytrasiewicz and Doshi 2005). I-POMDPs generalize POMDPs to multiagent settings by considering dynamic behavioral models of other agents as part of the state space. These models may themselves be I-POMDPs thereby leading to an infinitely-nested modeling space. In order to make the framework computable, the nesting is limited to a strategy level, $l$, thereby leading to finitely-nested I-POMDPs, which make the framework operational. Formally, a level $l$ I-POMDP for agent $i$ interacting with one other agent $j$ is defined as the following tuple:

$$\text{I-POMDP}_{i,l} = \langle IS_{i,l}, A, T_i, \Omega_i, O_i, R_i, OC_i \rangle$$

where:

- $IS_{i,l}$ denotes the set of *interactive states* at strategy level, $l$, defined as, $IS_{i,l} = S \times M_{j,l-1}$, where $S$ is the set of physical states, and $M_{j,l-1}$ is the set of models ascribed to the other agent. We describe the model space after this definition in this subsection.
- $A = A_i \times A_j$, is the set of joint actions of both agents.
- $T_i : S \times A \times S \to [0,1]$, is the transition function which gives the distribution over the next physical states given the current state and a joint action.
- $\Omega_i$ is the set of observations agent $i$ may receive.
- $O_i : A \times S \times \Omega_i \to [0,1]$, is the observation function which is the probability with which agent $i$ receives an observation conditioned on a joint action and the resulting state.
- $R_i : S \times A \to \mathbb{R}$, is the reward function which is the reward agent $i$ receives given a joint action performed by both agents from a state.
- $OC_i$ is the optimality criterion. In this article, we focus on optimizing the summed reward discounted over an infinite number of remaining steps, called the horizon.

Dennett's intentional stance (Dennett 1971) offers a way to organize the space of mental models into those that are *intentional* and denoted by $\Theta_j$, and others that are subintentional, denoted by $SM_j$. Intentional models ascribe beliefs, capabilities, preferences and rationality in action selection to the other agent. Examples of intentional models include the decision-theoretic formalism of POMDPs. Subintentional models include a distribution over actions, $\Delta(A_j)$, which may be history dependent as in fictitious play (Fudenberg 1998). Here, $\Delta(\cdot)$ denotes the set of all probability distributions over the argument random variable. A special example is the no-information model often represented by a uniform distribution. A more powerful subintentional model is the finite state automaton. We may follow a recursive bottom-up construction of the interactive state space.

$$\begin{aligned}
IS_{i,0} &= S, & \Theta_{j,0} &= \{\langle b_{j,0}, \hat{\theta}_{j,0}\rangle | b_{j,0} \in \Delta(IS_{j,0})\}, \\
& & M_{j,0} &= \Theta_{j,0} \cup SM_j \\
IS_{i,1} &= S \times M_{j,0}, & \Theta_{j,1} &= \{\langle b_{j,1}, \hat{\theta}_{j,1}\rangle | b_{j,1} \in \Delta(IS_{j,1})\}, \\
& & M_{j,1} &= \Theta_{j,1} \cup M_{j,0} \\
&\vdots & &\vdots \\
IS_{i,l} &= S \times M_{j,l-1}, & \Theta_{j,l} &= \{\langle b_{j,l}, \hat{\theta}_{j,l}\rangle | b_{j,l} \in \Delta(IS_{j,l})\}, \\
& & M_{j,l} &= \Theta_{j,l} \cup M_{j,l-1}
\end{aligned} \tag{1}$$

The 0-th level belief is a probability distribution over the physical states only, and the 0-th level models, $M_{j,0}$, are generally limited to be computable and consist of the set of computable intentional models of level 0, $\Theta_{j,0}$, and the subintentional models, $SM_j$. An intentional model, $\theta_{j,0} = \langle b_{j,0}, \hat{\theta}_{j,0}\rangle$, where $b_{j,0}$ is agent $j$'s level 0 belief, $b_{j,0} \in \Delta(IS_{j,0})$, and $\hat{\theta}_{j,0} = \langle A_j, T_j, \Omega_j, O_j, R_j, OC_j\rangle$, is collectively labeled as $j$'s level 0 *frame*. Here, $j$ is assumed to be Bayesian and rational. 0-th level intentional models are the traditional POMDPs whose parameters, $T_j$, $O_j$, and $R_j$ are specified over $j$'s individual actions, $A_j$. [1] [2] An agent's first-level beliefs are joint probability distributions over the physical states and level 0 models of the other agent. First-level models are computable and include computable level 1 intentional models and level 0 models of the agent. A level 1 intentional model, $\theta_{j,1} = \langle b_{j,1}, \hat{\theta}_{j,1}\rangle$, consists of the agent's first-level belief, $b_{j,1}$, and its frame, $\hat{\theta}_{j,1} = \langle A, T_j, \Omega_j, O_j, R_j, OC_j\rangle$. Note that parameters in the level 1 frame, $T_j$, $O_j$ and $R_j$ are specified over the joint actions, $A$. An agent's second-level beliefs are distributions over the physical states and level 1 models of the other agent, and so on up to the level $l$. In settings involving multiple agents, the actions of agent $j$ is replaced by the joint actions of all other agents and its model is replaced by the joint models of all other agents.

The agent bases its strategy on its current belief and picks the policy that would maximize its long turn expected reward. However, with each source of uncertainty, the complexity of solution increases making exact algorithms intractable. The uncertainties are captured in terms of the following curses:

- **Curse of dimensionality** results from the size of interactive state space.
- **Curse of history** is a result of the number of observations and gets exponentially worse between two consecutive time step.
- **Curse of recursive reasoning** is a an effect of curse of history on the model space of other agent and indirectly the dimensionality of the subject agents interactive state

[1] Note that the definition of a belief rests on first defining the underlying state space. The state space is not explicitly stated in the intentional model for brevity.

[2] One way of obtaining a POMDP at level 0 is to use a fixed distribution over the other agent's actions and fold it into $T_j$, $O_j$, and $R_j$ as noise.

space. The size of model space grows exponentially between two consecutive time step thereby further aggravating the curse of dimensionality.

- **Curse of many agents** manifests at each time step as the size of joint model space and joint action space grows exponentially with number of agents in the environment.

In my dissertation, I propose approaches to mitigate the effect of each of these curses.

## Current Research

In this section I describe the research that I have completed as a part of my PhD dissertation.

### Identifying Exploiting Weak Information Inducing Actions in Solving POMDPs

In this work (Sonu and Doshi 2011), I present a method for identifying actions that lead to observations which are only weakly informative in the context of partially observable Markov decision processes (POMDP). We call such actions as weak- (inclusive of zero-) information inducing. Policy subtrees rooted at these actions may be computed more efficiently. Specifically, I consider actions that lead to observations that tend to be only weakly informative. As an example, observations made during movement by a robotic vehicle (typically modeled sequentially post action in a POMDP) tend to be far less informative than those resulting from an action dedicated to observing. I call such actions as weak-information inducing; these include those that induce no information as well. I provide a simple and novel definition for weak information-inducing actions, characterizing the weakness of the observations using a parameter. Observing that policy trees rooted at zero-information inducing actions may be compressed, I utilize a simplified backup process that excludes considering observations for any weak-information inducing action while solving POMDPs. This results in significant computational savings, albeit we are currently unable to upper bound the error in optimality that this approximation introduces in the POMDP solution. I demonstrate the significant computational savings by exploiting such actions in the context of an exact solution technique incremental pruning (IP) (Cassandra et al. 1997) and the well-known point-based value iteration (PBVI) (Pineau et al. 2003), and empirically show that the solutions are of comparable quality. Ignoring weak information inducing actions mitigates the curse of history and the ideas presented in this paper may be extended to I-POMDPs.

### Generalized and Bounded Policy Iteration for Finitely Nested I-POMDPs

In this work (Sonu and Doshi 2012; 2014) I address the curses of dimensionality and recursive reasoning. One class of POMDP solution techniques deals with searching the policy space for the optimal policy. This class of techniques is known as *policy iteration*. I generalize policy iteration for POMDPs to I-POMDPs for significant scalability. The motivation behind this work is that while the model space of an agent it continuous, the models of the other agents may be grouped into discreet sets that are behaviorally equivalent thereby drastically reducing the dimensionality of the problem. I achieve such compression in dimensionality by representing the policy of the other agent as *finite state controllers* (FSCs) where each node of the controller represents a the root node of a policy and each edge represents transition on receiving an observation. This reduces the interactive state space to a discreet set. However, the size of such controllers could still grow exponentially between time steps (curse of recursive reasoning). To tackle this problem, I generalize bounded policy iteration (BPI) (Poupart et al. 2003) to I-POMDPs as interactive bounded policy iteration (I-BPI). BPI improves the current policy (controller) using a linear program approach that attempts to replace a node with a better valued node hence keeping the controller size fixed. The drawback of such an approach is that the controllers tend to converge to a local optima. However, local optima could be easily escape by adding few nodes so that the controller size doesn't explode as a result.

Drawing from the results of BPI, I first reformulate the interactive state space of level $l$ I-POMDP as the join of the physical state space and the nodes of a level $l-1$ I-POMDP controller. The policy iteration is carried in a bottom up fashion and is interleaved between levels to facilitate anytime solution. I-BPI outperforms and demonstrates significant scalability over the previous state of the art algorithm, interactive point based value iteration (I-PBVI) (Doshi and Perez 2008), in terms of the size of the problem (mitigating curse of dimensionality), and number of levels and frames of the agents (mitigating curse of recursive reasoning). Besides, for the first time we are able to solve problems involving multiple other agents.

### Bimodal Switching for Online Planning in Multiagent Settings

Next I further improve on addressing the curse of dimensionality in online settings (Sonu and Doshi 2013). I first prove that in multiagent settings where the observations of the subject agent are not affected by the actions of other agent but only depend on the physical state and the agent's own action, the observations are more informative when the entropy of the belief over physical states is low, i.e. when the agent is less uncertain about the current physical states. Based on this observation, I present a novel two-stage approach that focuses first on online planning as if the agent is alone, treating the other agents as noise, in order to reduce uncertainty in its belief over the physical state. In this mode, the agent is modeled as a POMDP and utilizes a fast POMDP-based planning technique, SARSOP (Kurniawati et al. 2008), that takes orders of magnitude less time to execute as compared to the I-POMDP solver. Subsequently, the agent switches to the I-POMDP model combining its updated belief over the state and the initial belief over the models. It now performs online planning using interactive particle filtering (Doshi and Gmytrasiewicz 2009).

A key question is when should the agent switch from the POMDP to the I-POMDP mode? In order to answer this, the agent at every step computes lower and upper bounds on the optimal decision at that step. The agent switches to the latter mode when the fractional difference between the lower and upper bounds at any step become less than a parameter, $\epsilon$. Because of the convexity property of the lower-bound value function (the POMDP described earlier), the difference between the two typically reduces as beliefs become less uncertain. The computational savings result because during the initial steps of online planning, a fast and scalable single-agent approach is utilized.

### Individual Planning in Agent Populations

Of all the curses that afflict I-POMDP solutions, the curse of many agents is the most severe. It affects not only the time complexity of the solution but also that of the memory required to represent the problem. The number of joint actions and joint models grows exponentially with the number of agents. Thus the storage requirement for the transition, observation, and reward function grows exponentially and so does the complexity of the iterative algorithms used to solve the I-POMDP. While the previous approaches for solving I-POMDPs focus on mitigating the curses of dimensionality, history, and recursive reasoning, there has been no directed effort to solve I-POMDPs involving many agents. As a result I-POMDPs haven't been solved for problems involving more than five agents and that too for the simplistic tiger problem (Gmytrasiewicz and Doshi 2005). In this work (Sonu et al. 2015) I exploit problem structures such as anonymity and context specific independence that are inherent in many real world problem domain and demonstrate the scalability to problems involving upto 1500 agents in a reasonable time.

In order to scale to a massive number of agents, I first formalize a factored representation of the agent's belief and derive a factored belief update for the same. Next, extending the work presented in *action graph games* (AGGs) (Jiang, Leyton-Brown, and Bhat 2011) in game theory literature, I propose utilizing an extensive version of anonymity and context-specific independence for dramatic scalability in I-POMDPs. In many settings, the transition, observation, and reward function do not depend directly on the identities of agent performing each action as is represented by a joint action, but on the number of agents performing each action which is a much cruder representation of other agents' action at each time step and belongs to a much smaller set of action counts compared to the set of joint actions. Moreover, given a context the effects on state, observation, and reward may depend on the counts of only a few actions rather than counts of all actions. This is known as context-specific independence and it further compresses the set of action counts. The contexts may also depend on the frame of the other agent performing an action. I utilize hypergraphs to capture the context-specific independence inherent in the problem. At each step, a dynamic program maps agents belief and context to a distribution over the relevant action counts which is then utlized instead of joint actions in solution. I am currently working on further refining the approach towards publication in a journal. I plan on submitting the extended article to the Journal of Artificial Intelligence Research before my defense.

### Realistic Simulation Testbed for Multiagent Decision Making

Yet another aspect of multiagent decision making that has been gaining ground in recent times is its application in real world scanario. The utilization of drones in the field of defense has opened up a huge avenue for deployment of autonomous agents. These agents may operate individually (as in I-POMDPs) or as a team (as in Dec-POMDPs) to perform certain reconnaissance tasks. Hence such decision theoretic frameworks could be used to solve for optimal policies that would guide the agents' behavior. I worked on developing a realistic simulation testbed for evaluating such policies to test the effectiveness of these agents before they could be deployed in real world. It is called the Georgia Testbed for Autonomous Control of vehicles (GaTAC). GaTAC provides a low-cost, open-source, and flexible distributed environment for realistically simulating the problem domains and evaluating solutions produced by multiagent decision making algorithms. For this project, we utilized an off-shelf open-source three dimensional flight simulator and added a communication module and an autonomous control module. The communication module enables instances of GaTAC to communicate with each other and the autonomous control module is used to control an agent (an autonomous unmanned aerial vehicle) either manually by a human operator or by using policies generated by decision theoretic frameworks. The source code for GaTAC is available under the open source AGPL licence agreement.

## Future Research

In the future, I wish to continue my research autonomous agents and artificial intelligence in a broader sense. In this section, I describe some of the possible research directions that I wish to undertake in future.

### Scalable Autonomous Planning

I wish to continue the research I undertook for my dissertation on scalable approaches for multiagent decision making. I wish to explore the problem of autonomous multiagent decision making in both individual and team settings. While much progress has been made in this regard, we are still far from their application to solve real-world problems. Exploiting problem structure such as anonymity and context-specific independence is a good way to move forward. Many other such structures exist in particular problem domains that remain to be explored.

### Applications of Multiagent Planning

Also, I wish to explore application of autonomous multiagent decison making in real-world scenarios. Some of the scenarios where decision theoretic agents may be deployed include but are not limited to the field of robotics, disaster management,self driving vehicles, law enforcement and security, defense, healthcare in providing assistance to the elderly, disabled or the sick, routing ground and air traffic, auctioning, and finance. Robot soccer and other team sports require deployment of agents that could both collaborate with their teammates and outsmart their opponents. Neither Dec-POMDPs nor I-POMDPs capture such problem description completely by themselves. Such avenues remain to be explored.

### Multiagent Learning

Multiagent learning is yet another field that has garnered much interest over the years. Many times, the models of the other agents may not be available to the subject agent and need to be learnt based on their interactions. The learnt models may then be solved to compute an optimal policy for the subject agent. This field of research is particularly useful in ad-hoc settings where heterogeneous agents (agents designed by different teams) are required to interact with each other and in settings where agents may be dynamically added to or removed from the environment.

### Game Theory

I-POMDPs draw much inspiration from game theory (particularly bayesian games and stochastic games). Although game theory has been studied detail over many decades, multiplayer games are still largely unexplored and have a lot to offer. I wish work on stochastic games and bayesian games and explore their relation to I-POMDPs in the future.

## Expectations from Doctoral Consortium

I am very close to the completion of my dissertation and I believe that my research makes a strong thesis. After graduation, I seek to pursue a career in academic research. Currently I am looking for position as a post-doctoral research associate in a field close to my research area. Eventually, I plan on joining a tenure-track academic position. My paper titled ”Individual Planning in Agent Populations: Exploiting Anonymity and Frame-Action Hypergraphs” has been accepted for publication at ICAPS 2015 and I plan on attending the conference to present it. Through this doctoral consortium program I wish to interact with other doctoral students and mentors and present my research to them in hopes for possible future collaboration. I also seek advice from mentors regarding my career choice and regarding pertinent research that offer the most scope for personal growth. I would appreciate any critique that I may receive on my thesis so that I may improve it before completion.

## References

Bernstein, D; Givan, R; Immerman, N; and Zilberstein, S 2002. The complexity of decentralized control of Markov decision processes. In *Mathematics of operations research*, 819–840.

Cassandra, A; Littman, M; and Zhang, N. 1997. Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes. *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, 54–61.

Dennett, Daniel C. 1971. Intentional systems. *The Journal of Philosophy* 87–106.

Doshi, P., and Gmytrasiewicz, P. J. 2009. Monte Carlo sampling methods for approximating interactive POMDPs. *JAIR* 34:297–337.

Doshi, P., and Perez, D. 2008. Generalized point based value iteration for interactive POMDPs. In *AAAI*, 63–68.

Doshi, P.; Qu, X.; Goodie, A.; and Young, D. 2010. Modeling recursive reasoning in humans using empirically informed interactive POMDPs. In *AAMAS*, 1223–1230.

Gmytrasiewicz, P. J., and Doshi, P. 2005. A framework for sequential planning in multiagent settings. *JAIR* 24:49–79.

Fudenberg, Drew. 1998. The theory of learning in games. Volume 2, MIT Press.

Del Giudice, A; Gmytrasiewicz, P; and Bryan, J. 2009. Towards strategic kriegspiel play with opponent modeling. *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2* 1265–1266.

Jiang, A. X.; Leyton-Brown, K.; and Bhat, N. A. 2011. Action-graph games. *Games and Econ. Behavior* 71(1):141–173.

Kaelbling, L; Littman, M; Cassandra, A 1998. Planning and acting in partially observable stochastic domains. In *Artificial intelligence*, 99–134.

Koller, D., and Milch, B. 2001. Multi-agent influence diagrams for representing and solving games. In *IJCAI*, 1027–1034.

Kurniawati, H; Hsu, D; and Lee, W. 2008. SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In *Proc. Robotics: Science and Systems*.

Meissner, C. 2011. A complex game of cat and mouse. In *LLNL Science and Technology Review*, 18–21.

Ng, B.; Meyers, C.; Boakye, K.; and Nitao, J. 2010. Towards applying interactive POMDPs to real-world adversary modeling. In *IAAI*, 1814–1820.

Pineau, J; Gordon, G; Thrun, S; and others 2003. Point-based value iteration: An anytime algorithm for POMDPs. In *IJCAI* 1025–1032.

Poupart, P.; and Boutilier, Craig 2003. Bounded finite state controllers. In *Advances in neural information processing systems*.

Seymour, R., and Peterson, G. L. 2009. Responding to Sneaky Agents in Multi-agent Domains. In *FLAIRS Conference*.

Seymour, R., and Peterson, G. L. 2009. A trust-based multiagent system. In *IEEE ICCSE*, 109–116.

Sonu, E., and Doshi, P. 2011. Identifying and exploiting weak-information inducing actions in solving POMDPs. *10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011* 1259–1260.

Sonu, E., and Doshi, P. 2012. Generalized and bounded policy iteration for finitely-nested interactive POMDPs: scaling up. *11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012* 1039–1048.

Sonu, E.; and Doshi, P. 2013. Bimodal Switching for Online Planning in Multiagent Settings. *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* 360–366.

Sonu, E.; and Doshi, P. 2014. Scalable solutions of interactive POMDPs using generalized and bounded policy iteration. *JAAMAS* DOI: 10.1007/s10458–014–9261–5, 1-40.

Sonu, E.; Chen, Y; and Doshi, P. 2015. Individual Planning in Agent Populations: Exploiting Anonymity and Frame-Action Hypergraphs. *Proceedings of the Twenty-Fifth International Conference on Automated Planning and Scheduling*.

Wang, F. 2013. An I-POMDP based multi-agent architecture for dialogue tutoring. In *ICAICTE*, 486–489.

Woodward, M. P., and Wood, R. J. 2012. Learning from humans as an i-pomdp. *CoRR* abs/1204.0274.

Wunder, M.; Kaisers, M.; Yaros, J.; and Littman, M. 2011. Using iterated reasoning to predict opponent strategies. In *AAMAS*, 593–600.