Derivation of Minimal Mental Models

David V. Pynadath and Stacy C. Marsella
USC Information Sciences Institute,
4676 Admiralty Way,
Marina del Rey CA 90292 USA
{pynadath,marsella}@isi.edu

1 Introduction

A teacher deciding how to maintain discipline may find it useful to keep track of which students (dis)like each other. In general, enriching the mental models that the teacher has of her students enables her to make better decisions. On the other hand, it is harder for her to maintain correct beliefs over the richer models. Intuitively, we expect a diminishing return on enriching the mental models, where adding more details offers less gain in accuracy in beliefs and less benefit in decision-making quality, while incurring additional overhead in maintaining those beliefs. For example, while the teacher could also keep track of her students' musical performances, she would expect little benefit to doing so. In contrast a student may expect considerable benefit in keeping track of other student's musical interests.

This basic issue of forming and maintaining models of others is not unique to human social interaction. Agents in general face the challenge of forming and updating their mental models of each other in a wide range of multiagent domains. Research in plan recognition has produced an array of techniques for modeling a planning agent and forming a belief about what its goals and intentions are, so as to predict its future actions [4, 6]. User modeling faces a similar problem in trying to understand and anticipate the needs of human users interacting with a software system [2]. Agents working together as teams must maintain beliefs about their teammates' status [3]. Social simulation of human social behavior may require agents with a theory of mind about the other agents in their society [5]. In games of incomplete information, each player faces uncertainty about the payoffs that the other players will receive [1].

In these domains, forming mental models is typically treated as a separate subproblem outside the decision-making context of the agent. The modeling agent starts from an initial set of possible models for the other agents, whether in the form of plan libraries in plan recognition, possible mental models in social simulation, private types in games of incomplete information, etc. As the modeling agent interacts with the other agents, it updates that belief based on its observations of their behavior. The modeling agent then uses its mental models of the other agents to make informed decisions based on expectations of what they will do.

In this paper, we observe that we can quantify the tradeoff by taking the problem of modeling others out of its isolation and placing it back within the overall decision-making context of the modeling agent. Doing so allows the agent to automatically derive a space of mental models according to an informed analysis of the cost-benefit tradeoffs.

Our approach comprises three methods: *Behavior equivalence*, where the modeling agent clusters models that lead to the same behaviors in its decision-making context; *Utility equivalence*, where the modeling agent clusters models that may lead to different behaviors, but produce equally preferred outcomes with respect to its utility; and *Approximate Utility Equivalence*, where the modeling agent clusters models that lead to performance losses that are below a certain threshold, sacrificing a fixed amount of accuracy.

We envision several benefits from these approaches. In most multiagent domains, agents can expect that this analysis will allow them to drastically reduce the original full mental model space, without overly sacrificing performance. Additionally, in simulation research on human social interaction, it establishes a normative baseline for the simplifications and distortions in people's mental models of others or theory of mind.

2 Modeling Other Agents

Across the various multiagent domains already mentioned (and even within each domain itself) researchers have applied a wide variety of possible modeling frameworks. We present a methodology using an abstract agent framework that is general enough to cover these approaches, as well as other decision-making procedures in the literature. When applying our methodology to a specific domain, these components would become specialized to the particular framework used for the agents in that domain.

2.1 Agent Notation

In general, an agent consists of its beliefs (including those about other agents), its actions, and its preferences. We use the same structure to represent both the actual agents and the mental models they have of each other. Thus, we represent the multiagent system as a set of real agents, $\{m_i\}_{i=1}^{N}$. Each such agent includes possible beliefs over mental models, M_{ij}, that represent what agent i can think of agent j. The modeling agent wishes to minimize this space, M_{ij}. In particular, we want an algorithm that computes the expected utility derived by modeling agent i when using the set of mental model spaces, $\{M_{ij}\}_{j=1}^{N}$, for all of the agents j in the system. We define the behavior of an agent as a policy, $\pi : B \rightarrow A$, out of a set of possible policies, Π. Any agent architecture will include an algorithm for translating an agent into such a policy, π. We will abstract this procedure into a generic function SOLVE: $M \rightarrow \Pi$, that takes an agent model (whether real or subjective) and returns that model's policy of behavior.

2.2 Example Domain

We have taken our example domain from a scenario in childhood aggression, modeled within PsychSim, a multiagent social simulation tool [5]. There are agents for three students: a bully, his victim (i.e., the student he focuses his aggression on), and an onlooking student to whom the bully looks for affirmation. There is also a teacher who can deter the bully from picking on his victim by doling out punishment. We focus on the problem facing the bully agent, whose decision on whether or not to pick on his victim must consider the possible punishment policy of the teacher.

2.2.1 Utility

PsychSim uses a decision-theoretic model of preferences, so the bully agent decides whether to pick on his victim through maximization of his utility, which has three components: (1) a desire to increase his power, which decreases when he is punished; (2) a desire for affirmation from the onlooking student, which increases when the onlooker laughs along; and (3) a desire to decrease the victim's power, which decreases when the bully picks on him (as well as when the onlooker laughs at him). The bully's utility function is a linear combination of these three components, so that we specify his type as a triple of coefficients, each in $[0, 1]$. Thus, to simulate the behavior of a bully whose aggression is intended to gain the approval of his peers, we would use an agent with a higher weight for the second component. On the other hand, to simulate a more sadistic bully, we would use a higher weight for the third. The teacher's utility also has three components, corresponding to her desire to increase the power of each of the three students. She thus has a disincentive for punishing anyone unless doing so will deter acts that would reduce the victim's power even more. A fair teacher would give equal weight to the three students' power. A bully feeling persecuted by the teacher may think that she favors the victim's power over his own. On the other hand, a bully may feel that the teacher shares his dislike of the victim, in which case he may model her as having a lower weight for the victim. We focus on the bully's modeling of the teacher, so we fix the onlooker to value his power (i.e., he does not want to be punished), while also wanting to decrease the victim's power out of dislike (i.e., he enjoys laughing at the victim when the bully picks on him).

2.2.2 Actions

The teacher has 7 options in her action set, A_T. She can do nothing; she can scold the bully, onlooker, or the entire class; or she can punish the bully, onlooker, or the entire class. Punishing a student causes a more severe decrease in a student's power than simply scolding. The onlooking student has 2 options in his action set, A_O: laugh at the victim, or do nothing. The bully has 2 actions in his action set, A_B: pick on the victim or do nothing.

2.2.3 Policies

To reduce the domain to its most essential, the bully's policy, $\pi_B : M_{BO} \times M_{BT} \rightarrow A_B$, is a function of his mental model of the onlooker and teacher. Given that the onlooker has only one possible mental model, the policy space for the bully, Π_B, contains $|A_B|^{|M_{BT}|}$ distinct policies. Thus, the complexity of the bully's problem of choosing his correct policy is highly dependent on the number of mental models that he must consider for the teacher. Similarly, the onlooker's policy, $\pi_O : M_{OB} \times M_{OT} \rightarrow A_O$, depends on only his mental model of the bully and the teacher. In this current investigation, we focus on only one entry in π_O, namely the one where $m_{OB} = m_B$ and $m_{OT} = m_{BT}$, where there are only two possible values: laughing at the victim or not. We must also specify what the bully expects the teacher to do, which depends on not only her mental models of the students, but also on the prior actions of the students ($\pi_T : M_{TB} \times M_{TO} \times A_B \times A_O \rightarrow A_T$). In other words, the teacher may perform a different action when the bully picks on the victim than when he does not. The bully assumes that the teacher knows the correct model of him (i.e., $m_{TB} = m_B$) and shares his mental model of the onlooker (i.e., $m_{TO} = m_{BO}$). Even with our simplifications, there still remains a large space of possible behaviors for the teacher: $|\Pi_T| = |A_T|^{|A_B| \cdot |A_O|} = 2401$.

2.2.4 Solution Mechanism

We use boundedly rational agents, so the bully's SOLVE algorithm performs a forward projection over his possible actions and chooses the action with the highest expected utility. The forward projection includes the bully's action, the onlooker's subsequent response, and the teacher's resulting punishment decision. To determine the teacher's policy, the bully applies a SOLVE method from the teacher's perspective that exhaustively tries all policies in Π_T, computes the best-response policies for the bully and onlooker, and then chooses the best policy based on her expected utility. Given the teacher's policy, the bully and onlooker can then choose their best-response policies. We can specify the bully's mental model of the teacher in terms of the three utility weights that the bully attributes to her. In other words, our initial space of possible mental models, M_{BT}, contains one model for every vector of weights, $\vec{w} = [w_B, w_O, w_V]$. For the purposes of this paper we discretize this space to contain the vectors $[0.0, 0.0, 1.0]$, $[0.0, 0.1, 0.9]$, $[0.0, 0.2, 0.8]$, $\ldots$, $[1.0, 0.0, 0.0]$, with a total size of 66 possible mental models that the bully can have of the teacher (i.e., $|M_{BT}| = 66$). The bully agent's decisions are highly dependent on what he expects the teacher to do. For example, if he picks on the victim, he is more likely to be severely punished by a teacher for whom the victim is a pet (i.e., for which w_V is high), but he would be more likely to escape punishment if he himself is a favorite of the teacher (i.e., if w_B is high). Thus, there is clearly some value to be gained by maintaining differential mental models of the teacher. However, from a psychological point of view, it is unlikely that real-life bullies juggle 66 possible mental models of their teachers in their heads, so the space is a good candidate for reduction.

This scenario is illustrative, and there are clearly many dimensions along which we could enrich it. For example, we could introduce state dependencies (e.g., the weaker the victim, the more damage done by picking on him). However, while these additional wrinkles would change the particular answers provided by our methodology, they would not change the *ability* of the methods presented in the following sections to provide such answers. Our core methodology presents a very general approach to quantifying the value of different mental model spaces even in the face of these additional complications. Therefore, we have removed as many extraneous domain features as possible, so as to be able to provide the clearest illustration of the methods and how they can be applied to any multiagent domain.

3 Behavior Equivalence

The modeling agent's goal is to find a minimal set of mental models that it needs to consider for the other agents. In looking for possible bases for such minimization, we observe that the modeling agent's

decisions often depend on only the *behavior* of the agents being modeled. Agents model the hidden parameters of others so as to generate expectations of their resulting behavior, but given the behavior of others, an agent's decision making is conditionally independent of the parameters behind it. For example, in agent teamwork, the mental states of the individual members have no direct effect on performance; only the decisions (actions, messages, etc.) derived from those mental states matter. Similarly, in games, the payoffs received by the agents depend on only the moves chosen by the players. In social simulations, the agents cannot read each others' minds, so they can base their decisions on only their observable behaviors. Therefore, regardless of what underlying parameters govern the modeled agent's decision-making, its eventual behavior is what has an impact on the modeling agent.

3.1 Behavior Equivalence Algorithm

This observation forms the basis for our first method for reducing the space of mental models. If two mental models produce the same behavior for the modeled agent, then making a distinction between them does not help the modeling agent. Therefore, it can safely remove one of them from consideration. It can do so by computing the policies corresponding to the possible mental models and clustering all that generate the same policy. The modeling agent then chooses one representative model from each cluster and removes all other models in the cluster from the overall space.

Algorithm 1 BEHAVIOREQUIVALENCE(M)

1: **for all** $m_1 \in M$ **do**
2: **for all** $m_2 \in M, m_1 \neq m_2$ **do**
3: **if** SOLVE($m1$) = SOLVE($m2$) **then**
4: remove m_2 from M

For many domains, the repeated invocations of the SOLVE function can be computationally intensive, but there is plenty of opportunity for specialization of Algorithm 1. For example, if the mental models correspond to points in a utility space (as in our social simulation domain), it should be possible to compare mental models to only their immediate neighbors. Furthermore, even if specializing the algorithm is insufficient, there are many opportunities for approximation as well. For example, one could easily re-write the loops in Lines 1 and 2 to implement a sampling algorithm that compares randomly selected pairs for behavior equivalence.

3.2 Behavior Equivalence Results

The bully agent starts with 66 possible mental models for the teacher in M_{BT}. It can apply behavior equivalence to reduce the size of that set, but the policy chosen by the teacher also depends on her model of the bully. For example, different bullies may be more afraid of a teacher punishing the whole class because of him than of being punished by himself. We thus performed a behavior equivalence reduction of the mental model space across different types of bullies. To do so, we discretized the space of possible (real) bullies in the same way that we discretized the space of possible mental models of the teacher. Thus, we represent different types of bullies by different vectors of utility weights, $\vec{w} = [w_B, w_O, w_V]$, and discretize the set of possible types into 66 distinct such vectors, $[0.0, 0.0, 1.0]$, $[0.0, 0.1, 0.9]$, $[0.0, 0.2, 0.8]$, ..., $[1.0, 0.0, 0.0]$. Each of the 66 possible bully types started with an initial space, M_{BT}, of the 66 possible mental models for the teacher. We gave the teacher and onlooker the correct model of the bully and of each other. 8 types of bullies reduced the number of mental models of the teacher from 66 to 4. The other 58 types of bullies reduced the number of mental models of the teacher from 66 to 5. Behavior equivalence provides a clear benefit to these bully agents. In particular, it is notable that, although the 66 types of teachers had 2401 policies to choose from, a specific bully could expect to come across only 4 or 5 distinguishable teacher behaviors. In fact, looking across the results for all of the possible bully types, there were only 8 policies that were *ever* selected by the teacher in the $66 \cdot 66 = 4356$ bully-teacher combinations. The reason that so much of the teacher's policy space is undesirable for her is that the bully's behavior is constrained by his utility. For example, regardless of where in our utility space he is, the bully always prefers not being punished to being punished. Therefore, it would never make sense for the teacher to adopt a policy of punishing the bully if he does nothing to the victim and doing nothing to him if he does.

4 Utility Equivalence

There are some multiagent domains where the modeling agent derives some direct utility from the values of the intrinsic parameters. For example, in our social simulation, the teacher may prefer being liked by her students, rather than feared, even if both cases produce complete obedience. In such cases, using behavior equivalence may over-prune the mental model space. However, it is still safe to assume that the modeled agent matters only in so far as it affects the modeling agent's expected utility. The modeling agent is thus completely indifferent between different mental models that produce the same expected utility in its own execution.

4.1 Utility Equivalence Algorithm

This observation leads to our second method for reducing the mental model space. If the modeling agent does not lose any expected utility when using a particular mental model when the correct model is actually another, then distinguishing between the two does not help. Therefore, the modeling agent can compute its expected utility derived based on the policies corresponding to each of the possible mental models (of the modeled) and clustering all of the models that generate the same value when mistaken for each other. It then again chooses one representative model from each cluster.

Algorithm 2 UTILITYEQUIVALENCE(m, M)

1: **for all** $m_1 \in M$ **do**
2: **for all** $m_2 \in M, m_1 \neq m_2$ **do**
3: $\pi_1 \leftarrow$ SOLVE($m1$), $\pi_2 \leftarrow$ SOLVE($m2$)
4: $u_{\text{right}} \leftarrow EU[\text{SOLVE}(m|m_2)|\pi_2]$
5: $u_{\text{wrong}} \leftarrow EU[\text{SOLVE}(m|m_2)|\pi_1]$
6: **if** $u_{\text{wrong}} - u_{\text{right}} \leq 0$ **then**
7: remove m_2 from M

While behavioral equivalence requires only the modeled agent's policy, utility equivalence requires the further computation of the modeling agent's own best response to that policy. Line 5 shows that the modeling agent computes the expected utility (u_{wrong}) it will derive if it solves for its policy assuming that the modeled agent is of type m_2, when it is actually of type m_1. Line 4 computes its expected utility (u_{right}) when using that same policy when m_2 is the correct mental model. If the first is no lower than the second, then

the agent can feel free to use m_1 in place of m_2. Line 6 accounts for the possibility that the utility loss might actually be negative when the agent being modeled, in turn, has an incorrect model of the modeling agent. Over time, if the agent being modeled updates its belief about the modeling agent, then such a utility gain is unlikely, because the modeled agent could eventually settle on a best response to the modeling agent's misconception. However, in the transient behavior, the modeled and modeling agents may inadvertently act in ways that improve the modeling agent's utility, despite the error in mental models.

Algorithm 2 adds another round of calls to the SOLVE function beyond what behavioral equivalence requires. The additional cost comes with the benefit of lossless reduction of the mental model space that sacrifices no utility to do so.

4.2 Utility Equivalence Results

To cluster the bully's mental models of the teacher according to utility equivalence, we followed the same experimental setup as for behavior equivalence. The 66 types of bully agents ran Algorithm 2, starting with the full space of mental models, M_{BT}. For this scenario, behavior equivalence implies utility equivalence, as the bully derives no direct utility from the teacher's intrinsic parameters. We can thus cluster the utility equivalence results according to the further reductions in mental model space achieved from M^b_{BT}. Of the 58 bully types with $\left|M^b_{BT}\right| = 5$, 11 types of bullies reduced the number of mental models of the teacher from 66 to 2, while the other 47 types reduced the number of mental models of the teacher from 66 to 4. Of the remaining 8 bully types with $\left|M^b_{BT}\right| = 4$, all of them reduced the number of mental models of the teacher from 66 to 3. Furthermore, for every type of bully, the mental model spaces reduced by utility equivalence (denoted M^u_{BT}) are all strict subsets of those reduced by behavior equivalence.

Some of the clustering occurs for bullies with extreme utility weights. For example, to a bully who cares about only hurting the victim (i.e, $\vec{w} = [0.0, 0.0, 1.0]$), mental models that differ on whether he himself gets punished are equivalent, because he does not care about the decrease in his own power. However, mental models that differ on whether or not the *onlooker* gets punished are not equivalent, because he desires the onlooker to laugh at the victim as well, to maximize the damage inflicted on the victim's power. Some of the clustering in this experiment arises when using an incorrect mental model of the teacher increases the bully's expected utility. For example, two mental models of the teacher may differ regarding whether punishment of the onlooker. From the bully's point of view, if the onlooker laughs regardless of the teacher's policy, then the bully does not care whether the onlooker is punished. Thus, while these two mental models produce different teacher behaviors, they produce the same expected utility to the bully, who is then justified in ignoring the distinction between them.

5 Approximate Utility Equivalence

The reduction of mental model spaces according to utility equivalence is lossless with respect to the modeling agent's decision making. Any further clustering of mental models will cost the modeling agent utility. However, the modeling agent can reduce its cost of maintaining beliefs over the mental model space by also clustering those models together that sacrifice a small amount of utility.

5.1 Approximate Utility Equivalence Algorithm

This observation leads to our third method for reducing the space of possible mental models. We can easily adapt Algorithm 2 to be tolerant of any utility loss below some positive threshold.

Algorithm 3 UTILITYAPPROX(m, M, θ)

```
1: for all m_1 ∈ M do
2:   for all m_2 ∈ M, m_1 ≠ m_2 do
3:     π_1 ← SOLVE(m1), π_2 ← SOLVE(m2)
4:     u_right ← EU[SOLVE(m|m_2)|π_2]
5:     u_wrong ← EU[SOLVE(m|m_2)|π_1]
6:     if u_wrong − u_right ≤ θ then
7:       remove m_2 from M
```

This approximate algorithm is no more complex than that for utility equivalence. In fact, we can perform a reduction using utility equivalence by passing in a threshold $\theta = 0$ to Algorithm 3.

The pseudocode in Algorithm 3 is written to support execution with a fixed threshold in mind. Alternatively, one could perform Lines 1–5 and *then* choose an appropriate threshold, θ, to reduce the space to an appropriate size. In other words, one would first profile the possible errors that would be derived from incorrect mental models before choosing a clustering. One could also easily vary the computation to use error measures other than expected utility. For example, one might be interested in worst-case utility loss instead of expected-case. Simply replacing the expectation in Lines 4 and 5 with a maximization would make the desired adjustment. There are any number of variations that would similarly modify the optimality criterion used in weighing the utility lost from the mistaken mental model.

5.2 Approximate Utility Equivalence Results

Figure 1 shows the results across our three methods for mental model space reduction. Each path from left to right represents the size of the mental model space for at least one possible type of bully as we raise its tolerance for utility loss. At the y-axis, all of the bully agents have the original mental model space of size 66. Then we see that these agents can reduce that size to either 4 or 5 models, using only the behavior equivalence method. The next point shows that the bully agents have spaces of 3–5 mental models when using only the utility equivalence method. Continuing along a path to the right represents the further reduction in the mental model space that comes with clustering mental models that cost less than the given threshold of expected utility.

As another example, there are 7 bully types that follow a path that leads to a mental model space of size one with only 10% loss of expected utility. If bully agents of this type are willing to tolerate a small utility loss, they can do away with modeling the teacher altogether! At the opposite end of the spectrum, there is one bully type that follows the upper envelope of the graph. For this bully type, utility equivalence allows for a mental model space of size 4, down from the size 5 of the space using only behavior equivalence. However, we see that even if the bully is willing to tolerate a loss of 25% of its expected utility, it still needs this full space of 4 models. If it wants to reduce its mental model space by even only one element, it can incur an up to 50% loss in expected utility if it is wrong. This bully type is also one of 14 in our sample space for which even tolerating 100% utility loss is not sufficient to warrant eliminating mental modeling together. In these cases, using the wrong mental model will lead to

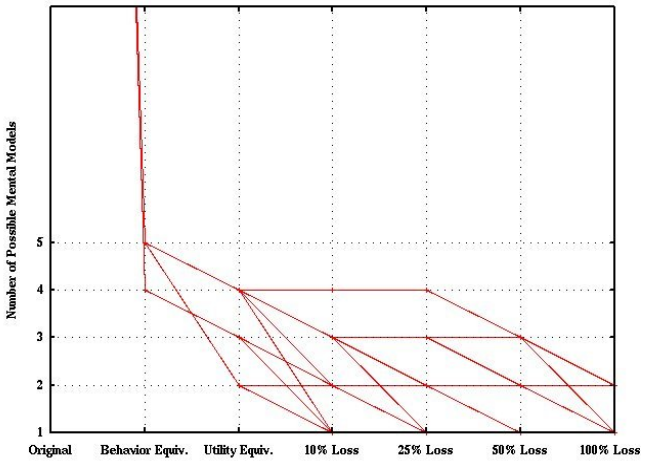

Figure 1. Size of model spaces vs. increasing leniency for utility loss, across all types of bully agents.

negative utility, so the bully has a strong incentive to do at least binary modeling of the teacher.

6 Discussion

While the exact graph in Figure 1 is specific to our example domain, it provides a concrete demonstration of our general ability to quantify the value of mental models to the modeling agent. To make the final decision, the agent must consider the computed value of the mental model space along with the cost of performing the actual model update and decision making during execution. As already described, the policy space of a modeling agent can grow exponentially with the number of mental models to consider. Furthermore, although we did not include the model update subproblem in our experiments, in most real-world domains, its complexity is highly dependent on the size of the mental model space. For example, probabilistic approaches, which compute a distribution over the possible mental models, can have a time complexity that is exponential in the size of the space. By finding minimal mental model spaces, an agent can apply more accurate belief update techniques that would have been computationally infeasible on larger spaces.

This methodology can also potentially create more psychologically plausible social simulations. In our experiments, the bully agents who were more attention-seeking (i.e., higher w_O) derived less value from the more complete mental model spaces for the teacher. Our characterization of bully types is consistent with the psychological literature that one can characterize different types of childhood aggression by the different goals that bullies have [7]. Thus, we can use our algorithms to explore the mental model spaces that we derive from those different goals and validate them against experimental data. Having validated the agents against such data, we can generate more confidence in the realism of the simulation.

We can also apply our algorithms to larger and more complicated domains. For example, our experiments have so far investigated the case of one agent choosing a space of mental models for only one other. Most multiagent domains will have multiple agents creating mental models of all of the others in the system. While our general methodology still applies in such cases, the additional interdependencies may lead to instabilities (e.g., an agent may be able to use a reduced space of mental models of another without utility loss only if the other uses a reduced space of mental models of him in return). Equilibrium concepts would provide one possible solution, but it may also be possible to re-cast our algorithms to simultaneously consider mental model spaces over the entire multiagent system, rather than over one modeling agent at a time.

While we deliberately designed this paper's domain to be simple enough to support a clear exposition and demonstration, we hope to learn more about the impact of design choices in mental modeling spaces and algorithms when we extend the analysis to richer domains. To support such domains, we will most likely have to implement some of the specialization and approximation techniques suggested in this paper. Once in place, we would be able to draw additional general conclusions about the impact of mental modeling choices as a function of fundamental properties of the multiagent system, and we expect that such general relationships may emerge on a richer class of domains.

7 Conclusion

At a higher level, the result of this investigation provides a key insight into the impact of social interaction on the design of multiagent systems. As designers, our immediate reaction is to view such interactions as complicating the problem of deriving appropriate multiagent behavior. However, as our results show, the interplay between the decision-making and modeling efforts of the individual agents is also highly *constraining* on that behavior. For example, out of the 2401 possible policies for the teacher, only 8 were ever desirable when interacting with our 66 types of bullies. When we view the problem of modeling other agents through the subjective lens of the modeling agent's own decision-making, we gain a utility metric that we can use both to restrict the scope of the modeling problem and to derive algorithms to solve it.

We used this metric to design algorithms that can quantify the value of distinctions made within the space of possible mental model space, and that then reduce that space accordingly. An agent can also use this same metric to derive a mental model space from scratch, simply by quantifying the value of *adding* mental models to the space of consideration. In this manner, our metric allows an agent designer to isolate those aspects of the mental models that are most relevant to the agent. We expect the algorithms to give such designers novel insight into the nature of their domains and to minimize the computational complexity of modeling other agents in all multiagent domains where such modeling is beneficial.

REFERENCES

[1] D. Fudenberg and J. Tirole, *Game Theory*, MIT Press, 1991.
[2] Anthony Jameson, 'Numerical uncertainty management in user and student modeling: An overview of systems and issues', *User Modeling and User-Adapted Interaction*, **5**(3-4), 193–251, (1995).
[3] Gal Kaminka, David V. Pynadath, and Milind Tambe, 'Monitoring teams by overhearing: A multi-agent plan-recognition approach', *Journal of Artificial Intelligence Research*, **17**, 83–135, (2002).
[4] Henry A. Kautz and James F. Allen, 'Generalized plan recognition', in *Proceedings of the National Conference on Artificial Intelligence*, pp. 32–37, (1986).
[5] David V. Pynadath and Stacy C. Marsella, 'PsychSim: Modeling theory of mind with decision-theoretic agents', in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1181–1186, (2005).
[6] C.F. Schmidt, N.S. Sridharan, and J.L. Goodson, 'The plan recognition problem: An intersection of psychology and artificial intelligence', *Artificial Intelligence*, **11**, 45–83, (1978).
[7] David Schwartz, 'Subtypes of victims and aggressors in children's peer groups', *Journal of Abnormal Child Psychology*, **28**, 181–192, (2000).