Modeling influence and theory of mind

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

Abstract

Agent-based modeling of human social behavior is an increasingly important research area. For example, such modeling is critical in the design of virtual humans, human-like autonomous agents that interact with people in virtual worlds. A key factor in human social interaction is our beliefs about others, in particular a theory of mind. Whether we believe a message depends not only on its content but also on our model of the communicator. The actions we take are influenced by how we believe others will react. In this paper, we present PsychSim, an implemented multiagent-based simulation tool for modeling interactions and influence among groups or individuals. Each agent has its own decision-theoretic model of the world, including beliefs about its environment and recursive models of other agents. Having thus given the agents a theory of mind, PsychSim also provides them with a psychologically motivated mechanism for updating their beliefs in response to actions and messages of others. We discuss PsychSim's architecture and its application to a school violence scenario.

1 Introduction

Modeling of human social behavior is an increasingly important research area (Liebrand et al. (1998)). A key factor in human social interaction is our beliefs about others, a *theory of mind* (Whiten (1991)). Specifically, the decisions we make and the actions we take are influenced by how we believe others will react. Similarly, whether we believe a message depends not only on its content but also on our model of the communicator.

Modeling theory of mind can play a key role in enriching social simulations. For example, childhood aggression is rooted in misattribution of another child's intent or outcome expectancies on how people will react to the violence (Schwartz (2000)). To develop a better understanding of the causes and remedies of school bullying, we can use agent models of the students that incorporate a theory of mind to simulate and study classroom social interactions. Models of social interaction have also been used to create social training environments where the learner explores high-stress social interactions in the safety of a virtual world (Marsella et al. (2000); Paiva et al. (2004)).

To facilitate such research and applications, we have developed a social simulation tool, called *PsychSim*, designed to explore how individuals and groups interact. PsychSim allows an end-user to quickly construct a social scenario, where a diverse set of entities, either groups or individuals, interact and communicate among themselves. Each entity has its own goals, relationships (e.g., friendship, hostility, authority) with other entities, private beliefs and mental models about other entities. The simulation tool generates the behavior for these entities and provides explanations of the result in terms of each entity's goals and beliefs. The richness of the entity models allows one to explore the potential consequences of minor variations on the scenario. A user can play different roles by specifying actions or messages for any entity to perform. Alternatively, the simulation itself can perturb the scenario to provide a range of possible behaviors that can identify critical sensitivities of the behavior to deviations (e.g., modified goals, relationships, or mental models).

A central aspect of the PsychSim design is that agents have decision-theoretic models of others. Such quantitative recursive models give PsychSim a powerful mechanism to model a range of factors in a principled way. For instance, we exploit this recursive modeling to allow agents to form complex attributions about others, enrich their messages to include the beliefs and goals of other agents, model the impact such recursive models have on an agent's own behavior, model the influence that observations of another's behavior have on the agent's model of that other, and enrich the explanations provided to the user. The decision-theoretic models in particular give

our agents the ability to judge degree of credibility of messages in a subjective fashion that factors in a range of influences that sway such judgments in humans. In this paper, we present PsychSim and discuss key aspects of its approach to modeling social interaction, specifically how people's actions and communications influence the beliefs and behaviors of others.

2 PsychSim Overview

PsychSim allows the setup of individuals or groups in a social environment and the exploration of how those entities interact. It has been designed to be a general, flexible multi-agent simulation tool.[1] The user sets up a simulation in PsychSim by selecting generic agent models that will play the roles of the various groups or individuals to be simulated and specializing those models as needed. To facilitate setup, PsychSim uses an automated fitting algorithm. For example, if the user wants the bully to initially attack a victim and wants the teacher to threaten the bully with punishment, then the user specifies those behaviors and the model parameters are fitted accordingly (Pynadath and Marsella (2004)). This degree of automation significantly simplifies simulation setup.

Execution of the simulation allows one to explore multiple tactics for dealing with a social issue and to see potential consequences of those tactics. How might a bully respond to admonishments, appeals to kindness or punishment? How might other groups react in turn? What are the predictions or unintended side-effects?

Finally, there is an analysis/perturbation capability that supports the iterative refinement of the simulation. The intermediate results of the simulation (e.g., the reasoning of the agents in their decision-making, their expectations about other agents) are all placed into a database. Inference rules analyze this database to explain the results to the user in terms of the agents' motivations, including how their beliefs and expectations about other agents influenced their own behavior and whether those expectations were violated. Based on this analysis, the system also reports sensitivities in the results, as well as potentially interesting perturbations to the scenario.

The rest of this paper describes PsychSim's underlying architecture in more detail, using a school bully scenario for illustration. The agents represent different people and groups in the school setting. The user can analyze the simulated behavior of the students to explore the causes and cures for school violence. One agent represents a bully, and another represents the student who is the target of the bully's violence (for young boys, the norm would be physical violence, while young girls tend to employ verbal abuse and ostracizing). A third agent represents the group of onlookers, who encourage the bully's exploits by, for example, laughing at the victim as he is beaten up. A final agent represents the class's teacher trying to maintain control of the classroom, for example by doling out punishment in response to the violence.

[1] For example, PsychSim is used in the Tactical Language simulation-based language training environment. The learner is immersed in an virtual facsimile of a foreign country, populated with animated characters that can talk to the learner in their native tongue. The characters are PsychSim agents.

3 The Agent Models

We embed PsychSim's agents within a decision-theoretic framework for quantitative modeling of multiple agents. Each agent maintains independent beliefs about the world, has its own goals and it owns policies for achieving those goals. The PsychSim framework is an extension to the Com-MTDP model (Pynadath and Tambe (2002)). Com-MTDP operates under the assumption that the agents are a member of a team. Therefore, to extend the Com-MTDP framework to social scenarios (where the agents are pursuing their own goals, rather than those of a team), we had to design novel agent models for handling belief update and policy application. This section describes the various components of the resulting model.

3.1 Model of the World

Each agent model starts with a representation of its current state and the Markovian process by which that state evolves over time in response to the actions performed by all of the agents.

State: Each agent model includes several features representing its "true" state. This state consists of objective facts about the world, some of which may be hidden from the agent itself. For our example bully domain, we included such state features as `power(agent)`, to represent the strength of an agent, though the agent may have its own subjective view of its own power. It is impacted by acts of violence, conditional on the relative powers of the interactants. `trust(truster,trustee)` represents the degree of trust that the agent `truster` has in another agent `trustee`'s messages. `support(supporter,supportee)` is the strength of support that an agent `supporter` has for another agent `supportee`. We represent the

state as a vector, $\vec{s}^t$, where each component corresponds to one of these state features and has a value in the range $[-1, 1]$.

Actions: Agents have a set of actions that they can choose to perform in order to change the world. An action consists of an action type (e.g., `punish`), an agent performing the action (i.e., the actor), and possibly another agent who is the object of the action. For example, the action `laugh(onlooker, victim)` represents the laughter of the onlooker directed at the victim.

World Dynamics: The state of the world changes in response to the actions performed by the agents. We model these dynamics using a transition probability function, $T(\vec{s}, \vec{a}, \vec{s}')$, to capture the possibly uncertain effects of these actions on the subsequent state:

$$\Pr(\vec{s}^{t+1} = \vec{s}' | \vec{s}^t = \vec{s}, \vec{a}^t = \vec{a}) = T(\vec{s}, \vec{a}, \vec{s}') \quad (1)$$

For example, the bully's attack on the victim affects the power of both the bully and victim. The distribution over the changes in power is a function of the relative powers of the two—e.g., the larger the power gap that the bully enjoys over the victim, the more likely the victim is to suffer a big loss in power.

3.2 Goals

An agent's goals represent its incentives (and disincentives) for behavior. In PsychSim's decision-theoretic framework, we represent goals as a reward function that maps the current world into a real-valued evaluation of benefit. We separate components of this reward function into two types of subgoals. A goal of **Minimize/maximize** `feature(agent)` corresponds to a negative/positive reward proportional to the value of the given state feature. For example, an agent can have the goal of maximizing its own power. A goal of **Minimize/maximize** `action(actor, object)` corresponds to a negative/positive reward proportional to the number of matching actions performed. For example, the teacher may have the goal of minimizing the number of times any student teases any other.

We can represent the overall goals of an agent, as well as the relative priority among them, as a vector of weights, $\vec{g}$, so that the product, $\vec{g} \cdot \vec{s}^t$, quantifies the degree of satisfaction that the agent receives from the world, as represented by the state vector, $\vec{s}^t$. For example, in the school violence simulation, the bully's reward function consists of goals of maximizing `power(bully)`, minimizing `power(victim)`, minimizing `power(teacher)`, and maximizing `laugh(onlookers, victim)`. We can model a sadistic bully with a high weight on the goal of minimizing `power(victim)` and an attention-seeking bully with a high weight on maximizing `laugh(onlookers, victim)`. In other words, by modifying the weights on the different goals, we can alter the motivation of the agent and, thus, its behavior in the simulation.

3.3 Beliefs about Others

As described by Sections 3.1 and 3.2, the overall decision problem facing a single agent maps easily into a partially observable Markov decision problem (POMDP) (Smallwood and Sondik (1973)). Software agents can solve make such a decision using existing algorithms to form their beliefs and determine the action that maximizes their reward given those beliefs. However, we do not expect people to conform to such optimality in their behavior. Thus, we have taken the POMDP algorithms as our starting point and and modified them in a psychologically motivated manner to capture more human-like behavior. This "bounded rationality" better captures the reasoning of people in the real-world, as well providing the additional benefit of avoiding the computational complexity incurred by an assumption of perfect rationality.

3.3.1 Nested Beliefs

The simulation agents have only a *subjective* view of the world, where they form beliefs, denoted by the vector $\vec{b}^t$, about what they *think* is the state of the world, $\vec{s}^t$. Agent A's beliefs about agent B have the same structure as the real agent B. Thus, our agent belief models follow a recursive structure, similar to previous work on game-theoretic agents (Gmytrasiewicz and Durfee (1995)). Fortunately, although infinite nesting of these agent models is required for modeling optimal behavior in software agents, *people* rarely use such deep models (Taylor et al. (1996)). In our implementation of the school violence scenario, the real agents are 2-level agents. In other words, they model each other as 1-level agents, who, in turn, model each other as 0-level agents, who do *not* have any beliefs. Thus, there is an inherent loss of precision (but with a gain in computational efficiency) as we move deeper into the belief structure.

Thus, each agent's beliefs consist of models of all of the agents (including itself), representing their state, beliefs, goals, and policy of behavior. For example, an agent's beliefs may include its subjective view on states of the world: "The bully believes that the teacher is weak", "The onlookers believe that

the teacher supports the victim", or "The bully believes that he/she is powerful." These beliefs may also include its subjective view on beliefs of other agents: "The teacher believes that the bully believes the teacher to be weak." An agent may also have a subjective view of the *goals* of other agents: "The teacher believes that the bully has a goal to increase his power." It is important to note that we also separate an agent's subjective view of itself from the real agent. We can thus represent errors that the agent has in its view of itself (e.g., the bully believes himself to be stronger than he actually is).

Actions affect the beliefs of agents in several ways. For example, the bully's attack may alter the beliefs that agents have about the state of the world—such as beliefs about the bully's power. Each agent updates its beliefs according to its subjective beliefs about the world dynamics. It may also alter the beliefs about the bully's goals and policy. We discuss the procedure of belief update in Section 3.4.

3.3.2 Policies of Behavior

Each agent's policy is a function, $\pi(\vec{b})$, that represents the process by which it selects an action or message based on its beliefs. An agent's policy allows us to model critical psychological distinctions such as reactive vs. deliberative behavior. We model each agent's real policy as a bounded lookahead procedure that seeks to best achieve the agent's goals given its beliefs. To do so, the policy considers all of the possible actions/messages it has to choose from and measures the results by simulating the behavior of the other agents and the dynamics of the world in response to the selected action/message. Each agent i computes a quantitative value, $V_a(\vec{b}_i^t)$, of each possible action, a, given its beliefs, $\vec{b}_i^t$.

$$V_a(\vec{b}_i^t) = \vec{g}_i \cdot \vec{b}_i^t + \sum_{\vec{b}_i^{t+1}} V(\vec{b}_i^{t+1}) \cdot \Pr(\vec{b}_i^{t+1} | \vec{b}_i^t, a, \vec{\pi}_{\neg i}(b_i^t)) \quad (2)$$

$$V(\vec{b}_i^t) = \vec{g}_i \cdot \vec{b}_i^t + \sum_{\tau=1}^{N} \sum_{\vec{b}_i^{t+\tau}} \vec{g}_i \cdot \vec{b}_i^{t+\tau} \cdot \Pr(\vec{b}_i^{t+\tau} | \vec{b}_i^{t+\tau-1}, \vec{\pi}(b_i^{t+\tau-1})) \quad (3)$$

Thus, an agent first uses the transition function, T, to project the immediate effect of the action, a, and then projects another N steps into the future, weighing each state along the path against its goals, $\vec{g}$. At the first step, agent i uses its model of the policies of all of the other agents, $\vec{\pi}_{\neg i}$, and, in subsequent steps, it uses its model of the policies of all agents, including itself, $\vec{\pi}$. Thus, the agent is seeking to maximize the expected value of its behavior, along the lines of decision policies in decision theory and decision theory. However, PsychSim's agents are only boundedly rational, given that they are constrained, both by the finite horizon, N, of their lookahead and the possible error in their belief state, $\vec{b}$.

3.3.3 Stereotypical Mental Models

If we applied this full lookahead policy in all of the nested models of the other agents, the computational complexity of the overall lookahead would quickly become infeasible as the number of agents grew. To simplify the agents' reasoning, we implement these mental models as simplified stereotypes of the richer lookahead models. For our simulation model of a bullying scenario, we have implemented mental models corresponding to *selfishness*, *altruism*, *dominance-seeking*, etc. For example, a model of a selfish agent specifies a goal of increasing its power as paramount, while a model of an altruistic agent specifies a goal of helping the weak. Similarly, a model of an agent seeking dominance specifies a goal of having relatively more power than its competitors.

These simplified mental models also include potentially erroneous beliefs about the policies of other agents. In particular, although the real agents use lookahead exclusively when choosing their own actions (as described in Section 3.3.2), the agents *believe* that the other agents follow much more reactive policies as part of their mental models of each other. PsychSim models reactive policies as a table of "Condition⇒Action" rules.

These more reactive policies in the mental models that agents have of each other achieves two desirable results. First, from a human modeling perspective, the agents perform a shallower reasoning when thinking about other agents, which more closely matches the shallower reasoning that people in the real world do of each other. Second, from a computational perspective, the direct action rules are cheap to execute, so the agents gain significant efficiency in their reasoning by avoiding expensive lookahead.

3.4 Influence and Belief Change

3.4.1 Messages

PsychSim views messages as attempts by one agent to influence the beliefs of recipients. Messages have five components: a source, recipients, a message subject, content and overhearers. For example, the teacher

(source) could send a message to the bully (recipient) that the principal (subject of the message) will punish acts of violence by the bully (content). Finally, overhearers are agents who hear the message even though they are not one of the intended recipients. Messages can refer to beliefs, goals, policies, or any other aspect of other agents. Thus, a message may make a claim about a state feature of the message subject ("the principal is powerful"), the beliefs of the message subject ("the principal believes that he is powerful"), the goals of the message subject ("the bully wants to increase his power"), the policy of the message subject ("if the bully thinks the victim is weak, he will pick on him"), or the stereotypical model of the message subject ("the bully is selfish").

3.4.2 Influence Factors

A challenge in creating a social simulation is addressing how groups or individuals influence each other, how they update their beliefs and alter behavior based on observations of, as well as messages from, others. Although many psychological results and theories must inform the modeling of such influence (e.g., Cialdini (2001); Abelson et al. (1968); Petty and Cacioppo (1986)), they often suffer from two shortcomings from a computational perspective. First, they identify factors that affect influence but do not operationalize those factors. Second, they are rarely comprehensive and do not address the details of how various factors relate to each other or can be composed. To provide a sufficient basis for our computational models, our approach has been to distill key psychological factors and map those factors into our simulation framework. Here, our decision-theoretic models are helpful in quantifying the impact of factors and in such a way that they can be composed.

Specifically, a survey of the social psychology literature identified the following key factors:

Consistency: People expect, prefer and are driven to maintain consistency, and avoid cognitive dissonance, between beliefs and behaviors. This includes consistency between their old and new information, between beliefs and behavior, as well as consistency with the norms of their social group.

Self-interest: Self-interest impacts how information influences us in numerous ways. It impacts how we interpret appeals to one's self-interest, values and promises of reward or punishment. The inferences we draw are biased by self-interest (e.g., motivated inference) and how deeply we analyze information in general is biased by self-interest. Self-interest may be in respect to satisfying specific goals like "making money" or more abstract goals such as psychological reactance, the tendency for people to react to potential restrictions on freedom such as their freedom of choice (e.g., the child who is sleepy but refuses to go to bed when ordered by a parent.)

Speaker's Self-interest: If the sender of a message benefits greatly if the recipient believes it, there is often a tendency to be more critical and for influence to fail.

Trust, Likability, Affinity: The relation to the source of the message, whether we trust, like or have some group affinity for him, all impact whether we are influenced by the message.

3.4.3 Computational Model of Influence

To model such factors in the simulation, one could specify them exogenously and make them explicit, user-specified factors for a message. This tactic is often employed in social simulations where massive numbers of simpler, often identical, agents are used to explore emergent social properties. However, providing each agent with a model of itself and, more importantly, fully specified models of other agents gives us a powerful mechanism to model this range of factors in a principled way. We model these factors by a few simple mechanisms in the simulation: *consistency*, *self-interest*, and *bias*. We can render each as a quantitative function on beliefs that allows an agent to compare alternate candidate belief states (e.g., an agent's original $\vec{b}$ vs. the $\vec{b}'$ implied by a message).

Consistency is an evaluation of whether the content of a message or an observation was consistent with prior observations. In effect, the agent asks itself, "If this message is true, would it better explain the past better than my current beliefs?". We use a Bayesian definition of consistency based on the relative likelihood of past observations given the two candidate sets of beliefs (i.e., my current beliefs with and without believing the message). An agent assesses the quality of the competing explanations by a re-simulation of the past history. In other words, it starts at time 0 with the two worlds implied by the two candidate sets of beliefs, projects each world forward up to the current point of time, and compares the projected behavior against the behavior it actually observed. In particular, the consistency of a sequence of observed actions, $\omega^0, \omega^1, \ldots$, with a given belief state, $\vec{b}$, corresponds to:

$$\begin{aligned} & \text{consistency}(\vec{b}^t, [\omega^0, \omega^1, \ldots, \omega^{t-1}]) \\ = \; & \Pr\left([\omega^0, \omega^1, \ldots, \omega^{t-1}] \,\middle|\, \vec{b}^t\right) \\ \propto \; & \sum_{\tau=0}^{t-1} V_{\omega^\tau}(\vec{b}^\tau) \end{aligned} \tag{4}$$

Thus, it must verify that the action that it thinks each agent would perform matches the action taken during the actual simulation. Note that the value function, V, computed is with respect to the agent performing the action at time τ. In other words, we are summing the value of the observed action to the acting agent, given the set of beliefs under consideration. The higher the value, the more likely that agent is to have chosen the observed action, and, thus, the higher the degree of consistency.

Self-interest is similar to consistency, in that the agent compares two sets of beliefs, one which accepts the message and one which rejects it. However, while consistency requires evaluation of the past, we compute self-interest by evaluating the future using Equation 3. An agent can perform an analogous computation using its beliefs about the sender's goals to compute the sender's self-interest in sending the message.

Bias factors act as tie-breakers when consistency and self-interest fail to decide acceptance/rejection. We treat support (or affinity) and trust as such a bias on message acceptance. Agents compute their support and trust levels as a running history of their past interactions. In particular, one agent increases (decreases) its trust in another, when the second sends a message that the first decides to accept (reject). This current mechanism is very simple, but our future work will explore the impact of using richer algorithms from the literature. Regarding changes in support, an agent increases (decreases) its support for another, when the second selects an action that has a high (low) reward, with respect to the goals of the first. In other words, if an agent selects an action a, then the other agents modify their support level for that agent by a value proportional to $\vec{g} \cdot \vec{b}$, where $\vec{g}$ corresponds to the goals and $\vec{b}$ the new beliefs of the agent modifying its support.

Upon receiving any information (whether message or observation), an agent must consider all of these various factors in deciding whether to accept it and how to alter its beliefs (including its mental models of the other agents). For a message, the agent determines acceptance using a weighted sum of the five components: consistency, self-interest, speaker self-interest, trust and support. For an observed action by an agent, all of the other agents first check whether the action is consistent with their current beliefs (including mental models) of that agent. If so, no belief change is necessary. If not, the agents evaluate alternate mental models as possible new beliefs to adopt in light of this inconsistent behavior. The other agents evaluate the possible belief changes using the same weighted sum as for messages, except that the speaker, in this case, is the agent about whom they are considering changing mental models.

In addition, each agent considers this belief update when doing its lookahead. In particular, Equations 2 and 3 project the future beliefs of the other agents in response to an agent's selected action. Thus, the agent's decision-making procedure is sensitive to the different effects each candidate action may have on the beliefs of others. Similar to work by de Rosis et al. (2003), this mechanism provides PsychSim agents with a potential incentive to deceive, if doing so leads the other agents to perform actions that lead to a better state for the deceiving agent.

We see the computation of these factors as a toolkit for the user to explore the system's behavior under existing theories that we can encode in PsychSim. For example, the elaboration likelihood model (ELM) (Petty and Cacioppo (1986)) argues that the way messages are processed differs according to the relevance of the message to the receiver. High relevance or importance would lead to a deeper assessment of the message, which is consistent with the self-interest calculations our model performs. For less relevant messages, more peripheral processing of perceptual cues such as "liking for" the speaker would dominate. PsychSim's linear combination of factors is roughly in keeping with ELM because self-interest values of high magnitude would tend to dominate. One could also realize non-linear combinations where this dominance of one factor over the other was more dramatic.

We could extend the use of these basic mechanisms to a range of phenomena. An agent could exploit his theory of mind to reason not only about consistency with respect to his beliefs, observations and models of others but also evaluate consistency with respect to special subclasses of beliefs (e.g., norms, values, cherished beliefs and ingroup vs. outgroup). Reactance/restriction of freedom could be in agent's reward function and therefore be factored into interest calculations. For example, in the School domain, the bully might have a reactance goal of not doing what it is told to do.

4 Example Scenario Operation

The research literature on childhood bullying and aggression provides interesting insight into the role that theory of mind plays in human behavior. Although a number of factors are related to bullying, two social cognitive variables have been shown to play a central role. One variable discussed is a hostile attributional style Nasby et al. (1979), wherein typical

playground behaviors are interpreted as having a hostile intent. Children who tend to see other children as intending to hurt them are more likely to display angry, retaliatory aggression. A second variable is outcome expectancies for the effectiveness of aggression. Children develop outcome expectancies for the effectiveness of aggression depending on whether in the past they have been rewarded for its use or found it to be ineffective or punished for it.

Investigations of bullying and victimization Schwartz (2000) have identified four types of children: those who characteristically display reactive aggression (aggressive victims), those who display proactive aggression (nonvictimized aggressors), those who are typically victimized (nonaggressive victims), and normal children. Nonaggressive victims display a hostile attributional style and have negative outcome expectancies for aggression. Aggressive victims tend to have a hostile attributional style, but neither positive nor negative outcome expectancies for aggression. Nonvictimized aggressors have positive outcome expectancies for aggression, but do not have a hostile attributional style.

We have begun to use PsychSim to explore psychological theories by demonstrating how PsychSim can represent both *attributional style* and *outcome expectancies* in a simulation of school violence. The user can manipulate each factor to generate a space of possible student behaviors for use in simulation and experimentation. For example, an agent's attributional style corresponds to the way in which it updates its beliefs about others to explain their behavior. A hostile attributional style corresponds to an agent who tends to adopt negative mental models of other agents. In our example scenario, agents with a hostile attributional style mentally model another student as having the goal of hurting them (i.e., minimizing their power).

Our agents already compute the second factor of interest, outcome expectancies, as the expected value of actions (V_a from Equation 2). Thus, when considering possible aggression, the agents consider the immediate effect of an act of violence, as well as the possible consequences, including the change in the beliefs of the other agents. In our example scenario, a bully has two incentives to perform an act of aggression: (1) to change the power dynamic in the class (i.e., weaken his victim and make himself stronger), and (2) to earn the approval of his peers (as demonstrated by their response of laughter at the victim). Our bully agent models the first incentive as a goal of maximizing `power(bully)` and minimizing `power(victim)`, as well as a belief that an act of aggression will increase the former and decrease the latter. The second incentive must also consider the actions that the other agents may take in response. The agents' theory of mind is crucial here, because it allows our bully agent to predict these responses, albeit limited by its subjective view.

For example, a bully motivated by the approval of his classmates would use his mental model of them to predict whether they would enjoy his act of aggression and laugh along with him. Similarly, the bully would use his mental model of the teacher to predict whether he will be punished or not. The agent will weigh the effect of these subjective predictions along with the immediate effect of the act of aggression itself to determine an overall expected outcome. Thus, the agents' ability to perform bounded lookahead easily supports a model for proactive aggression.

We explored the impact of different types of proactive aggression by varying the priority of the two goals (increasing power and gaining popularity) within our bully agent. When we ran PsychSim using an agent model where the bully cares about each goal equally, then the threat of punishment is insufficient to change the bully's behavior, because he expects to still come out ahead in terms of his popularity with his peers. On the other hand, a threat against the whole class in response to the bully's violence is effective, because the bully then believes that an act of violence will *decrease* his popularity among his peers. If we instead use an agent model where the bully favors the first goal, then even this threat against the whole class is ineffective, because the bully no longer cares about his popularity in the class.

Of course, this example illustrates one outcome, where we do not change any of the other variables (e.g., bully's power relative to victim, teacher's credibility of threats). PsychSim's full range of variables provide a rich space of possible class makeups that we can systematically explore to understand the social behavior that arises out of different configurations of student psychologies. We have also begun developing algorithms that are capable of finding the configuration that best matches a real-world class dynamic, allowing us to find an underlying psychological explanation for a specific instance of behavior (Pynadath and Marsella (2004)). Furthermore, as illustrated, we can try out different interventions in the simulation to understand their impact under varying student models. As we have seen, alternate scenarios will have different results, but by systematically varying the scenario, we can draw general conclusions about the effectiveness of these different intervention methods. Finally, although this section uses

a specific taxonomy of student behavior to illustrate PsychSim's operation, the methodology itself is general enough to support the exploration of many such taxonomies.

5 Conclusion

We have presented PsychSim, an environment for multi-agent simulation of human social interaction that employs a formal decision-theoretic approach using recursive models. Our agents can reason and simulate the behavior and beliefs of other agents with a theory of mind that allows them to communicate beliefs about other agent's beliefs, goals and intentions and be motivated to use communication to influence other agents' beliefs about agents. Within PsychSim, we have developed a range of technology to simplify the task of setting up the models, exploring the simulation, and analyzing results. This includes new algorithms for fitting multi-agent simulations. There is also an ontology for modeling communications about theory of mind. We have exploited the recursive models to provide a psychologically motivated computational model of how agents influence each other's beliefs. We believe PsychSim has a range of innovative applications, including computational social science and the model of social training environments. Our current goals are to expand the exploration already begun in the school violence scenario and begin evaluating the application of PsychSim there and in these other areas.

References

Robert P. Abelson, Eliot Aronson, William J. McGuire, T.M. Newcomb, M.J. Rosenberg, and Percy H. Tannenbaum, editors. *Theories of Cognitive Consistency: A Sourcebook*. Rand McNally, Chicago, IL, 1968.

Robert Cialdini. *Influence: Science and Practice*. Allyn and Bacon, Boston, MA, 2001.

Fiorella de Rosis, Cristiano Castelfranchi, Valeria Carofiglio, and Giuseppe Grassano. Can computers deliberately deceive? A simulation tool and its application to Turing's imitation game. *Computational Intelligence*, 19(3):253–263, 2003.

Piotr J. Gmytrasiewicz and Edmund H. Durfee. A rigorous, operational formalization of recursive modeling. In *Proceedings of the International Conference on Multi-Agent Systems*, pp. 125–132, 1995.

Wim Liebrand, Andrzej Nowak, and Rainer Hegselmann, editors. *Computer Modeling of Social Processes*. Sage, London, UK, 1998.

Stacy C. Marsella, W. Lewis Johnson, and Catherine LaBore. Interactive pedagogical drama. In *Proceedings of the International Conference on Autonomous Agents*, pp. 301–308, New York, 2000. ACM Press.

W. Nasby, B. Hayden, and B.M. DePaulo. Atttributional biases among aggressive boys to interpret unambiguous social stimuli as displays of hostility. *Journal of Abnormal Psychology*, 89:459–468, 1979.

Ana Paiva, Joao Dias, Daniel Sobral, Ruth Aylett, Polly Sobreperez, Sarah Woods, Carsten Zoll, and Lynne Hall. Caring for agents and agents that care: Building empathic relations with synthetic agents. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 194–201, New York, 2004. ACM Press.

Richard Petty and John Cacioppo. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer, New York, NY, 1986.

David V. Pynadath and Stacy C. Marsella. Fitting and compilation of multiagent models through piecewise linear functions. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 1197–1204, New York, 2004. ACM Press.

David V. Pynadath and Milind Tambe. Multiagent teamwork: Analyzing the optimality and complexity of key theories and models. In *Proceedings of the International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pp. 873–880, New York, 2002. ACM Press.

David Schwartz. Subtypes of victims and aggressors in children's peer groups. *Journal of Abnormal Child Psychology*, 28:181–192, 2000.

Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21:1071–1088, 1973.

Jasper Taylor, Jean Carletta, and Chris Mellish. Requirements for belief models in cooperative dialogue. *User Modelling and User-Adapted Interaction*, 6:23–68, 1996.

Andrew Whiten, editor. *Natural Theories of Mind*. Basil Blackwell, Oxford, UK, 1991.