Modeling Theory of Mind and Cognitive Appraisal with Decision-Theoretic Agents

David V. Pynadath[1], Mei Si[2], and Stacy C. Marsella[1]

[1]Institute for Creative Technologies, University of Southern California

12015 Waterfront Drive, Playa Vista, CA 90094-2536 USA

[2]Cognitive Science Department, Rensselaer Polytechnic Institute

110 8th Street, Troy, NY 12180 USA

pynadath@ict.usc.ed, sim@rpi.edu, marsella@ict.usc.edu

April 7, 2011

Abstract

Agent-based simulation of human social behavior has become increasingly important as a basic research tool to further our understanding of social behavior, as well as to create virtual social worlds used to both entertain and educate. A key factor in human social interaction is our beliefs about others as intentional agents, a *Theory of Mind*. How we act depends not only on the immediate effect of our actions but also on how we believe others will react. In this paper, we discuss PsychSim, an implemented multiagent-based simulation tool for modeling social interaction and influence. While typical approaches to such modeling have used first-order logic, PsychSim agents have their own decision-theoretic models of the world, including beliefs about their environment and recursive models of other agents. Using these quantitative models of uncertainty and preferences, we have translated existing psychological theories into a decision-theoretic semantics that allow the agents to reason about degrees of believability in a novel way. We demonstrate the expressiveness of PsychSim's decision-theoretic implementation of Theory of Mind by presenting its use as the foundation for a domain-independent model of appraisal theory, the leading psychological theory of emotion. The model of appraisal within PsychSim demonstrates the key role of a Theory of Mind capacity in appraisal and social emotions, as well as arguing for a uniform process for emotion and cognition.

1 Introduction

Computational models of social interaction have become increasingly important in both basic research on human social behavior and a range of applications. For example, computational models of psychological or sociological theories promise to transform how theories of human behavior are formulated and evaluated [43]. In addition, computational models of human social interaction have also become increasingly important as a means to create simulated social environments used in a variety of training and entertainment applications [21]. For example, many serious games have used models of social interaction as the basis for virtual (typically embodied) autonomous characters [2, 24, 33, 44, 46].

We argue that such models of social interaction must address the fact that people interact within a complex social framework. A central factor in social interaction is the beliefs we have about each other, a *Theory of Mind* [49]. Our choice of action is influenced by how we believe others will feel and react. Whether we believe what we are told depends not only on the content of the communication but also on our model of the communicator. How we emotionally react to another's action is influenced by our beliefs as well, for example whether we believe he or she intended to cause harm [30]. The central goal of our research is to bring such Theory of Mind capacities to the design of computational models of social interaction both as a basic research tool and as a framework for virtual character design.

Unfortunately, traditional artificial intelligence techniques are ill-suited for modeling Theory of Mind. Representations using first-order logic are often insensitive to the distinctions among conflicting goals that people must balance in a social interaction. For example, psychological research has identified a range of goals that motivate classroom bullies (e.g., peer approval, sadism, tangible rewards) [31]. Different bullies may share the same goals, but the relative priorities that they place on them will lead to variations in their behavior. Resolving the ambiguity among equally possible, but unequally plausible or preferred, options requires a quantitative model of uncertainty and preference. Unfortunately, more quantitative frameworks, like decision theory and game theory, face their own difficulties in modeling human psychology. Game-theoretic frameworks typically rely on concepts of equilibria that people rarely achieve in an unstructured social setting like a classroom. Decision-theoretic frameworks typically rely on assumptions of rationality that people violate.

We have developed a social simulation framework, PsychSim [20, 27], that operationalizes existing psychological theories as boundedly rational computations to generate more plausibly human behavior. PsychSim allows a user to quickly construct a social scenario where a diverse set of entities, groups or individuals, interact and communicate. Each entity has its own preferences, 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 preferences and beliefs. The richness of the entity models allows one to explore the potential consequences of minor variations on the scenario.

A central aspect of the PsychSim design is that agents have fully specified 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, send messages that include the beliefs and preferences of others, and use their observations of another's behavior to influence their model of that other.

In operationalizing psychological theories within PsychSim, we have taken a strong architectural stance. We assume that decision-theoretic agents that incorporate a Theory of Mind provide a uniform, sufficient computational core for modeling the factors relevant to human social interaction. While the sufficiency of our framework remains an open question, such a strong stance yields the benefit of uniform processes and representations that cover a range of phenomena. Our stance thus provides subsequent computational benefits, such as optimization and reuse of the core algorithms that provide the agent's decision-making and belief revision capacities.

More significantly, this uniformity begins to reveal common elements across apparently disparate psychological phenomena that typically have different methodological histories. To illustrate such common elements, we have demonstrated how a range of human psychological and social phenomena can be modeled within our framework, including wishful thinking [12], influence factors [20], childhood aggression [27] and emotion [39].

In this article, we discuss two of those models. First, we use a model of childhood aggression to motivate the discussion of the overall framework as well as to demonstrate its expressiveness. Second, in keeping with the theme of this volume, we go into considerable detail on how PsychSim's decision-theoretic agents with a Theory of Mind provide a particularly effective basis for a computational model of emotion.

Computational models of emotion have largely been based on appraisal theory [5, 7, 18, 22, 6, 23, 28, 47], a leading psychological theory of emotion. Appraisal theory argues that a person's subjective assessment of their relationship to the environment determines his or her emotional responses [8, 13, 14, 23, 29, 30, 41, 42]. This assessment occurs along several dimensions, such as motivational congruence, accountability, novelty and control. For example, an event that leads to a bad outcome for a person (motivationally incongruent) and is believed to be caused by others (accountability) is likely to elicit an anger response. On the other hand, if the event is believed to be caused by the person himself/herself, he/she is more likely to feel guilt or regret [29].

We approach the task of incorporating appraisal into the existing PsychSim multiagent framework as a form of thought experiment: Can we leverage the existing processes and representations in PsychSim to model appraisal? The motivations for this thought experiment are three-fold. First, we seek to demonstrate overlap between the theoretical

model of appraisal theory and decision-theoretic, social agents of PsychSim. Specifically, we are interested in whether appraisal offers a possible blueprint, or requirements specification, for intelligent social agents by showing that an existing framework not predesigned with emotion or appraisal in mind has, in fact, appraisal-like processes already present.

Conversely, we seek to illustrate the critical role that subjective beliefs about others plays in allowing agents to model social emotions. Because the agent's representations and decision-making processes are rooted in a Theory of Mind capacity, incorporating and maintaining beliefs about others, the appraisal process inherits this social frame, allowing the agent to appraise events from its own perspective as well as others. Thus, in keeping with the tenets of *social appraisal* [16], the behaviors, thoughts and emotions of the other can also be appraised and thereby influence the agent.

Finally, we seek a design that is elegant, by reusing architectural features to realize new capabilities such as emotion. Alternative approaches for creating embodied conversational agents and virtual agents often integrate separate modules for emotion, decision-making, dialogue, etc., which leads to sophisticated but complex architectures [44]. The work here is an alternative minimalist agenda for agent design. In particular, based on the core theory of mind reasoning processes, appraisal can be derived with few extensions.

We begin the paper with a demonstration of PsychSim's application to a childhood aggression scenario. We then discuss how PsychSim can represent appraisal theory and present a preliminary assessment of its implementation.

2 The Agent Models

This section describes PsychSim's underlying architecture, 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. 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. We embed PsychSim's agents within a decision-theoretic framework for quantitative modeling of multiple agents. Each agent maintains its independent beliefs about the world, has its own goals and it owns policies for achieving those goals.

2.1 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.

2.1.1 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. `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 $[-1, 1]$.

2.1.2 Actions

Agents have a set of actions that they can choose 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.

2.1.3 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, $T(\vec{s}_i, \vec{a}, \vec{s}_f)$, to capture the possibly uncertain effects of these actions on the subsequent state:

$$\Pr(\vec{s}^{t+1} = \vec{s}_f | \vec{s}^t = \vec{s}_i, \vec{a}^t = \vec{a}) = T(\vec{s}_i, \vec{a}, \vec{s}_f) \quad (1)$$

For example, the bully's attack on the victim impacts the power of the bully, the power of the victim, etc. The distribution over the bully's and victim's 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.

2.2 Preferences

PsychSim's decision-theoretic framework represents an agent's incentives for behavior as a reward function that maps the state of the world into a real-valued evaluation of benefit for the agent. 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 preferences 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)`, and maximizing `laugh(onlookers, victim)`. By modifying the weights on the different goals, we can alter the motivation of the agent and, thus, its behavior in the simulation.

2.3 Beliefs about Others

As described by Sections 2.1 and 2.2, the overall decision problem facing a single agent maps easily into a partially observable Markov decision problem (POMDP) [40]. Software agents can solve such a decision problem using existing algorithms to form their beliefs and then 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 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 as providing the additional benefit of avoiding the computational complexity incurred by an assumption of perfect rationality.

2.3.1 Nested Beliefs

The agents have only a *subjective* view of the world, where they form beliefs, $\vec{b}^t$, 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 [9]. Of course, the nesting of these agent models is potentially unbounded. However, although infinite nesting is required for modeling optimal behavior, people rarely use such deep models [45]. In our school violence scenario, we found that 2-level nesting was

sufficiently rich to generate the desired behavior. Thus, the agents 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.

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 *preferences* 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 preferences and policy. We discuss the procedure of belief update in Section 2.4.

2.3.2 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 maximize expected reward 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, $\vec{b}_i^t$.

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

$$V^N(\vec{b}_i^t) = \max_a V_a^N(\vec{b}_i^t) \tag{3}$$

The agent computes the posterior probability of subsequent belief states ($\Pr(\vec{b}_i^{t+1})$) by using the transition function, T, to project the immediate effect of the action, a, on its beliefs. It then projects another N steps into the future, weighing each state against its goals, $\vec{g}$. At the first step, agent i uses its model of the policies of all of the other agents, $\pi_{\neg i}$, and, in subsequent steps, it uses its model of the policies of all agents, including itself, π. Thus, the agent is seeking to maximize the expected reward of its behavior as in a POMDP. 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, $\vec{b}$. By varying N for different agents, we can model entities who display different degrees of reactive vs. deliberative behavior in their thinking.

2.3.3 Stereotypical Mental Models

If we applied this full lookahead policy within the nested models of the other agents, the computational complexity of the top-level lookahead would quickly become infeasible as the number of agents grew. To simplify the agents' reasoning, these mental models are realized as simplified stereotypes of the richer lookahead behavior models of the agents themselves. For our simulation model of a bullying scenario, we have implemented mental models corresponding to *attention-seeking*, *sadistic*, *dominance-seeking*, etc. For example, a model of an attention-seeking bully specifies a high priority on increasing the approval (i.e., `support`) that the other agents have for it, a dominance-seeking bully specifies a high priority on increasing its power as paramount, and a bully agent specifies a high priority on hurting others.

These simplified mental models also include potentially erroneous beliefs about the policies of other agents. Although the real agents use lookahead exclusively when choosing their own actions (as described in Section 2.3.2), 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. The left-hand side conditions may trigger on an *observation* of some action or a *belief* of some agent (e.g., the bully believing himself as powerful). The conditions may also be more complicated combinations of these basic triggers (e.g., a *conjunction* of conditions that matches when each and every individual condition matches).

The use of 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 that provides a more accurate model of the real-world entities they represent. Second, from a computational perspective, the direct action rules are cheap to execute, so the agents gain significant efficiency in their reasoning.

2.4 Modeling Influence and Belief Change

2.4.1 Messages

Messages are attempts by one agent to influence the beliefs of another. Messages have four components: source, recipients, subject, and content. For example, the teacher (source) could tell the bully (recipient) that the principal (subject of the message) will punish violence by the bully (content). Messages can refer to beliefs, preferences, policies, or any other aspect of other agents. Thus, a message may make a claim about a state feature of the subject

("the principal is powerful"), the beliefs of the subject ("the principal believes that he is powerful"), the preferences of the subject ("the bully wants to increase his power"), the policy of the subject ("if the bully thinks the victim is weak, he will pick on him"), or the stereotypical model of the subject ("the bully is selfish").

2.4.2 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 any partial observation of, as well as messages from, others. Although many psychological results and theories must inform the modeling of such influence (e.g., [1, 3, 25]) 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 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.

Self-interest: 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.

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.

2.4.3 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 quantitative models of itself and, more importantly, 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 of 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 the degree to which a potential belief agreed with prior observations. In effect, the agent asks itself, "If this belief holds, 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 (e.g., 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 computes the probability of the observation it received. The higher the value, the more likely that agent is to have chosen the observed action, and, thus, the higher the degree of consistency.

In previous work, we have investigated multiple methods of converting such action values into a degree of consistency [11]. For the purposes of the current work, we use only one of those methods, defining the consistency of a sequence of observations, $\omega^0, \omega^1, \ldots$, with a given belief state, $\vec{b}$, as follows:

$$\text{consistency}(\vec{b}^t, \left[\omega^0, \omega^1, \ldots, \omega^{t-1}\right]) = \Pr\left(\left[\omega^0, \omega^1, \ldots, \omega^{t-1}\right] \middle| \vec{b}^t\right) \propto \sum_{\tau=0}^{t-1} \sum_{a \in A} e^{\text{rank}(V_a(\vec{b}^\tau) \Pr(\omega^\tau | a, \vec{b}^\tau))} \quad (4)$$

The algorithm first ranks the utilities of the actor's alternative actions in reversed order (rank(v)). The value function, V, computed is with respect to the agent performing the action at time τ. Thus, the higher the rank of the likelihood of the observation, the more consistent it is with the candidate belief state.

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 evaluates 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 preferences to compute the sender's self-interest in sending the message.

Bias factors represent subjective views of the message sender that influence the receiver's acceptance/rejection of the message. 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). Similarly, an agent increases (decreases) its support for another, when the second selects an action that has a high (low) reward, with respect to the preferences 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. Whenever an agent observes an action by another, it checks whether the observation is consistent with its current beliefs (including mental models). If so, no belief change is necessary. If not, the agent evaluates alternate mental models as possible new beliefs to adopt in light of this inconsistent behavior. Agents evaluate these possible belief changes using the same weighted sum as for messages.

Each agent's decision-making procedure is sensitive to these changes that its actions may trigger in the beliefs of others. Each agent accounts for the others' belief update when doing its lookahead, as Equations 2 and 3 project the future beliefs of the other agents in response to an agent's selected action. Similar to work by [4] 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, which we can encode in PsychSim. For example, the elaboration likelihood model (ELM) [25] 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. PsychSim's linear combination of factors is roughly in keeping with ELM because self-interest values of high magnitude would tend to dominate.

3 Childhood Aggression Model

The research literature on childhood aggression provides interesting insight into the role that Theory of Mind plays in human behavior. Investigations of bullying and victimization [31] have identified four types of children; we focus here on *nonvictimized aggressors*, those who display proactive aggression due to positive outcome expectancies for aggression. Children develop expectations on the likely outcomes of aggression based on past experiences (e.g., did past acts of aggression lead to rewards or punishment). This section describes the results of our exploration of the space of different nonvictimized aggressors and the effectiveness of possible intervention strategies in dealing with them.

3.1 Scenario Setup

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. In our bullying scenario, we constructed generic bully models that compute 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 three subgoals that provide incentives to perform an act of aggression: (1) to change the power dynamic in the class by making himself stronger, (2) to change the power dynamic by weakening his victim, and (3) 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 the second as minimizing `power(victim)`, both coupled with a belief that an act of aggression will increase the former and decrease the latter. The third incentive seeks to maximize the `laugh` actions directed at the victim, so it must consider the actions that the other agents may take in response.

For example, a bully motivated by the approval of his classmates would use his mental model of them to predict whether they would laugh along with him. We implemented two possible mental models of the bully's classmates: *encouraging*, where the students will laugh at the victim, and *scared*, where the students will laugh only if the teacher did not punish them for laughing last time. Similarly, the bully would use his mental model of the teacher to predict whether he will be punished or not. We provide the bully with three possible mental models of the teacher: *normal*, where the teacher will punish the bully in response to an act of violence; *severe*, where the teacher will more harshly punish the bully than in the *normal* model; and *weak*, where the teacher never punishes the bully.

The relative priorities of these subgoals within the bully's overall reward function provide a large space of possible behavior. When creating a model of a specific bully, PsychSim uses a fitting algorithm to automatically determine the appropriate weights for these goals to match observed behavior. For example, if the user wants the bully to initially attack a victim and the teacher to threaten the bully with punishment, then the user specifies those behaviors and the model parameters are fitted accordingly [26]. This degree of automation significantly simplifies simulation setup. In this experiment, we selected three specific bully models from the overall space: (1) *dominance-seeking*, (2) *sadistic*, and (3) *attention-seeking*, each corresponding to a goal weighting that favors the corresponding subgoal.

3.2 Experimental Results

PsychSim allows one to explore multiple tactics for dealing with a social issue and see the potential consequences. Here, we examine a decision point for the teacher after the bully has attacked the victim, followed by laughter by the rest of the class. At this point, the teacher can punish the bully, punish the whole class (including the victim), or do nothing. We explore the impact of different types of proactive aggression by varying the type of the bully, the teacher's decision to punish the bully, the whole class, or no one, and the mental models that the bully has of the other students and the teacher.

A successful outcome is when the bully does not choose to act out violently toward the victim the next time around.

Bully Type	**Punish Whom?**	**Model of Students**	**Model of Teacher**	**Success?**
Sadistic	bully	*	¬severe	N
			severe	Y
	class	scared	*	N
		encouraging	¬severe	N
			severe	Y
	no one	*	¬severe	N
			severe	Y
Attention Seeking	bully	*	*	N
	class	scared	weak	N
			normal	Y
			severe	Y
		encouraging	*	N
	no one	*	*	N
Dominance Seeking	*	*	weak	N
			normal	Y
			severe	Y

Table 1: Outcomes of intervention strategies

By examining the outcomes under these combinations, we can see the effects of intervention over the space of possible classroom settings. Table 1 shows all of the outcomes, where we use the "*" wildcard symbol to collapse rows where the outcome was the same. Similarly, a row with "¬severe" in the **Teacher** row spans the cases where the bully's mental model of the teacher is either *normal* or *weak*.

We first see that the PsychSim bully agent meets our intuitive expectations. For example, we see from Table 1 that if the bully thinks that the teacher is too weak to ever punish, then no immediate action by the teacher will change the bully from picking on the victim. Thus, it is critical for the teacher to avoid behavior that leads the bully to form such mental models. Similarly, if the bully is of the *attention-seeking* variety, then punishment directed at solely himself will not dissuade him, as he will still expect to gain peer approval. In such cases, the teacher is better off punishing the whole class.

We can see more interesting cases as we delve deeper. For example, if we look at the case of a *sadistic* bully when the teacher punishes the whole class, we see that bully can be dissuaded only if he thinks that the other students will *approve* of his act of violence. This outcome may seem counter-intuitive at first, but the *sadistic* bully is primarily concerned with causing suffering for the victim, and thus does not mind being punished if the victim is punished as well. However, if the bully thinks that the rest of the class is *encouraging*, then the teacher's punishment of the whole class costs him peer approval. On the other hand, if the bully thinks that the rest of the class is already *scared*, so that they will not approve of his violence, then he has no peer approval to lose.

Such exploration can offer the user an understanding of the potential pitfalls in implementing an intervention

strategy. Rather than providing a simple prediction of whether a strategy will succeed or not, PsychSim maps out the key conditions, in terms of the bully's preferences and beliefs, on which a strategy's success depends. PsychSim provides a rich space of possible models that we can systematically explore to understand the social behavior that arises out of different configurations of student psychologies. We are continuing to investigate more class configurations and the effects of possible interventions as we expand our models to cover all of the factors in school aggression identified in the literature.

4 A Model of Appraisal

To further illustrate the capacity of these decision-theoretic agents to model social interaction, we also used them to implement a computational model of emotion, largely based on Smith and Lazarus's theory of cognitive appraisal [42]. As noted in Section 1, our implementation was driven by a thought experiment: Can we leverage the existing processes and representations in PsychSim to model appraisal? The motivations for this thought experiment are three-fold. First, we seek to demonstrate an intrinsic coupling between the theoretical model of appraisal and decision-theoretic social reasoning. Specifically, we wish to show that an existing social framework with no explicit emotion or appraisal capabilities has, in fact, appraisal-like processes already present as part of its decision-making processes. Second, we also seek to illustrate the critical role that subjective beliefs about others play in modeling social emotions. Finally, we seek a minimalist design that elegantly reuses architectural features to model new social phenomena, like emotion.

This work on modeling appraisal is in the spirit of EMA [10, 19] (see also Chapter X), which defines appraisal processes as operations over a plan-based representation (a *causal interpretation*) of the agent's goals and how events impact those goals. The agent's existing cognitive processes maintain the causal interpretation, which appraisal-specific processes leverage. Thus, in EMA, the cognitive processes for constructing the person-environment relation representation are distinct from appraisal itself, which is reduced to simple and fast pattern matching over the plan representation.

We seek to eliminate this distinction by demonstrating how the cognitive processes for decision-making can also generate appraisals. While EMA uses a uniform *representation* across cognition and appraisal, we seek to additionally establish uniformity over the *algorithms* underlying both cognition and appraisal. In so doing, we can identify that appraisal itself is already an integral part of the cognitive processes that a social agent must perform to maintain its beliefs about others and to inform its decision-making in a multiagent social context.

Specifically, we treat appraisal as leveraging component algorithms already present in a PsychSim agent by deriving key appraisal variables from the outputs of these algorithms. Furthermore, the appraisal process inherits the

intrinsic social nature of these algorithms, allowing the agent to appraise events from another's perspective, as well as from its own. Thus, in keeping with Manstead and Fischer's concept of social appraisal, the behaviors, thoughts and emotions of the other can also be appraised and thereby influence the agent.

We have modeled five appraisal dimensions so far: motivational relevance, motivational congruence, accountability, control and novelty. We adapted Smith and Lazarus's [42] definitions for modeling motivational relevance, motivational congruence and accountability. Our model of control is roughly equivalent to Smith and Lazarus's [42] definition of problem-focused coping potential. It is closer to Scherer's [30] definition of control, because it accounts for the overall changeability of the situation and not an individual agent's power to make a change. Finally, we modeled novelty based on Leventhal and Scherer's [15, 30] definition of predictability-based novelty, as there is no equivalent concept in Smith and Lazarus.

The model of appraisal is built within Thespian [33, 34, 35, 36, 37, 38, 39], which extends PsychSim for modeling and simulating computer-aided interactive narratives. This computational model of appraisal is one of Thespian's extensions. We demonstrate the application of the appraisal model in three different scenarios: a simple conversation between two people, a firing-squad scenario as modeled in [17], and a fairy tale, "The Little Red Riding Hood". The last scenario will be described in detail in the next section as it will be used as an example to motivate the discussion of our appraisal model. The details of the other two scenarios are given in Section 5.

4.1 Little Red Riding Hood Domain

The story contains four main characters, Little Red Riding Hood (Red), Granny, the hunter and the wolf. The story starts as Red and the wolf meet each other on the outskirts of a forest while Red is on her way to Granny's house. The wolf wants to eat Red, but it dares not because there is a wood-cutter close by. At this point, the wolf and Red can either have a conversation or go their separate ways. The wolf may have future chances to eat Red if he finds her alone at another location. Moreover, if the wolf hears about Granny from Red, it can even go to Granny's house and eat her as well. Meanwhile, the hunter is searching for the wolf to kill it. Once the wolf is killed, all of the wolf's previous victims can escape. Our model of this story builds upon the base PsychSim representation, as described in Section 2.

4.2 Modeling Appraisal in Thespian

Appraisal is a continuous process [42]. People constantly reevaluate their situations and cope with unfavorable situations, forming a "appraisal-coping-reappraisal" loop. In this section we illustrate how we can model this phenomenon and derive appraisal dimensions by leveraging algorithms and information within a PsychSim agent's belief revision and decision-making processes.

During decision making, the lookahead process generates the agent's expectations about possible future events so that it can choose the action with the most favorable outcome. Figure 1 is an example of one-step lookahead while assuming that the other characters will also perform a one-step lookahead. Bold shapes denote the actions with the highest expected utility based on the actor's mental models. When no shapes are in bold, there are multiple actions with the same expected utility, and the actor has no specific preference over which action will be picked. Finally, this example is a simplified version of our actual implementation, where the agents simulated more steps of lookahead and had more actions to choose from.

The agent remembers the expectations generated in its last two lookahead processes, because the evaluations of some appraisal dimensions (e.g. accountability) need to trace back more than one step. Note that these expectations contain not only the agent's expectations about its own and the other agents' future actions, but also the expected states/utilities of all possible action choices of each of the agents. These state/utility traces serve as required input for the explanation algorithms that provide the user with the agent's motivations for making its observed choices.

Upon observing a new event (an action performed by an agent or the user), each agent updates its beliefs and appraises the new situation. The calculation of motivational relevance, motivational congruence, novelty and accountability depends on only the agent's beliefs about the other agents' and its own utility values in the current and previous steps, which therefore can be derived immediately (see Section 4.3). Our calculation of control occurs only after the agent has made its own decision in response to the observed event. In fact, at this time the agent could potentially reevaluate every appraisal dimension as it computes updated information about expected states/utilities.

The mental models that Thespian agents have of each other enable them to not only have emotional responses to the environment but also form expectations of other agents' emotions. To simulate another agent's appraisal processes, the observing agent's beliefs about the other agent are used for deriving appraisal dimensions. For instance, the wolf can use its beliefs about Red to evaluate the motivational relevance and novelty of an event to her, which will most likely be totally different from the wolf's own evaluations of those dimensions. In our current implementation, if the observing agent has multiple mental models of other agents, currently it uses only the most likely mental models (rather than the entire distribution of likelihoods) to simulate their appraisals.

4.3 Appraisal Dimensions

This section illustrates how the agents can evaluate five appraisal dimensions (motivational relevance, motivation congruence or incongruence, accountability, control and novelty) using states/utilities calculated during the agents' belief revision and decision-making processes.

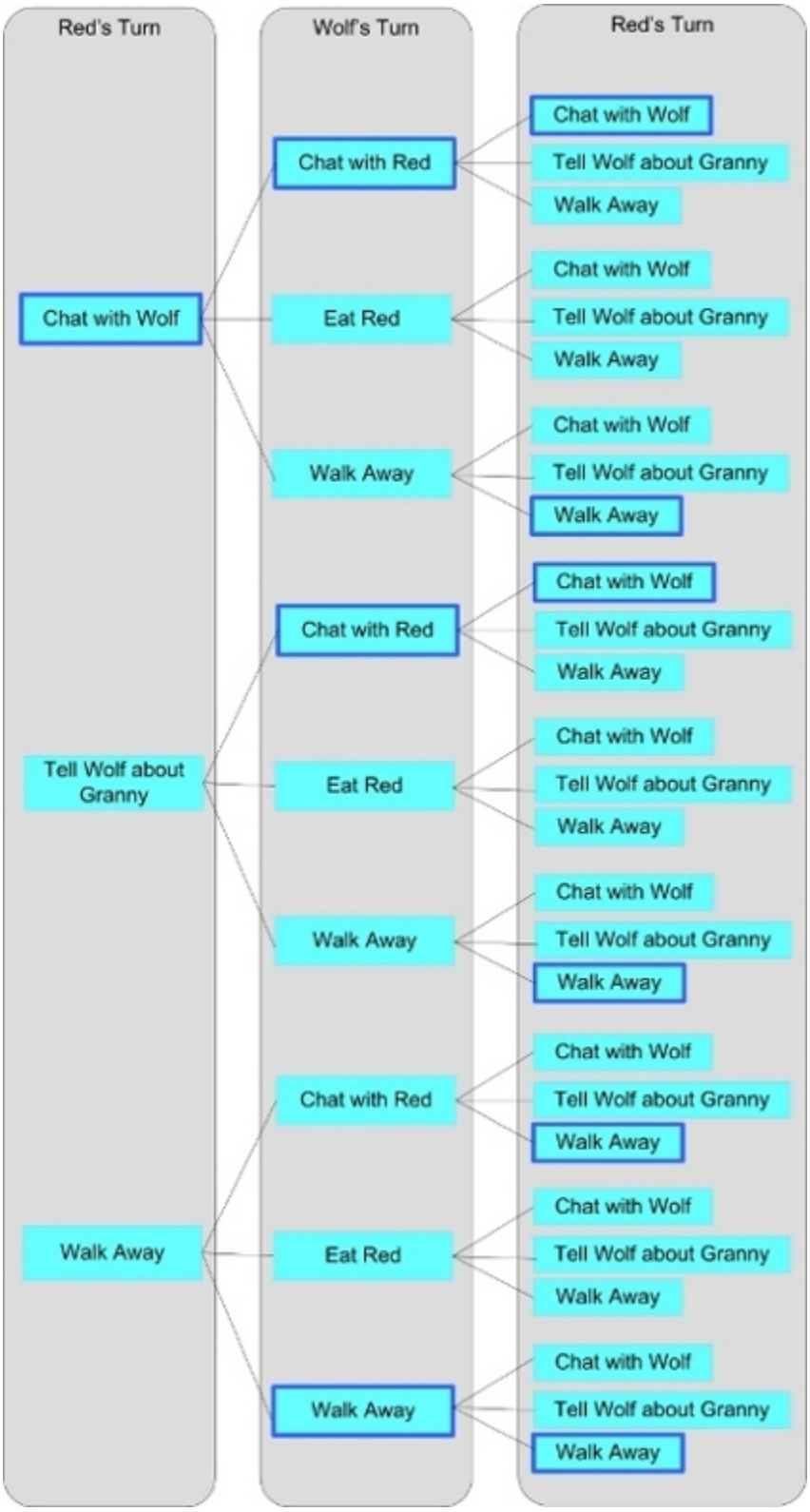

Figure 1: Red's lookahead process

4.3.1 Motivational Relevance & Motivational Congruence or Incongruence

Motivational relevance evaluates the extent to which an encounter touches upon personal goals. Motivational congruence or incongruence measures the extent to which the encounter thwarts or facilitates personal goals [42].

Algorithm 1 Motivational Relevance & Motivation Congruence

preUtility: utility before the event happens
curUtility: utility after the event happens

$$\textit{Motivational Relevance} = \left|\frac{\textit{curUtility}-\textit{preUtility}}{\textit{preUtility}}\right|$$

$$\textit{Motivational Congruence} = \frac{\textit{curUtility}-\textit{preUtility}}{|\textit{preUtility}|}$$

We model these appraisal dimensions as a product of the agent's utility calculations which are integral to the agent's decision-theoretic reasoning. We use the ratio of the relative utility change and the direction of the utility change to model these two appraisal dimensions. The rationale behind this is that the same amount of utility change will result in different subjective experiences depending on the agent's current utility. For instance, if eating a person increases the wolf's utility by 10, it will be 10 times more motivationally relevant and congruent when the wolf's original utility is 1 (very hungry) than when the wolf's original utility is 10 (less hungry).

Algorithm 1 evaluates motivational relevance and motivational congruence/incongruence. *preUtility* denotes the agent's expected utility before the other agent takes an action. For agents who perform at least one step of lookahead, this value is evaluated by the agents' previous decision-making process. Taking Figure 1 as an example, Red's expected utility is the sum of her utility over this sequence of actions: Red chats with the wolf, the wolf chats with Red and Red chats with the wolf again. This follows Red's expectations of the interaction. *curUtility* denotes the agent's updated expected utility, reflecting the effect of the other agent's action. Note that this value is also pre-computed when the agent generated its expectations about the other agent's possible actions. For example, if the wolf chooses the action as Red has expected, then *curUtility* is the same as *preUtility* for Red. If the wolf instead chooses to walk away, then Red's expectation about future events changes. In this case, she would expect herself to walk away too and *curUtility* would be the sum of Red's utilities over this alternate sequence of actions: Red chats with the wolf, the wolf walks away and Red walks away too. If an agent does not perform any lookahead (i.e., it cares about only its immediate reward and not the responses of the other agents), the value of *curUtility* will not be calculated by the agent's previous decision-making process. It *will* be calculated when the agent updates its beliefs to compute its immediate reward in the next time step, and the evaluation of the appraisal dimensions will happen at that time.

The sign of *Motivational Congruence* indicates whether the event is motivationally congruent or incongruent to the agent. When the value is negative, the event is motivationally incongruent to the extent of *Motivational Relevance*,

and otherwise the event is motivationally congruent to the agent.

Agents can also have goals to help or hinder other agents in the pursuit of their own goals. If an action helps advance a Thespian agent's self-centered goals but hurts a friend's goals (where a friend is someone whom the agent has a goal to help), then the agent's overall utility is diminished, subsequently muting the action's motivational congruence and relevance. For example, while Red will derive a self-centered satisfaction from eating the cake, that satisfaction will be offset by the negative impact on Granny, who has been deprived of the cake's benefit.

4.3.2 Accountability

Accountability determines who deserves credit or blame for a given event [42]. Various theories have been proposed for assigning blame/credit, e.g. [32, 48]. The reasoning usually considers factors such as who caused the event, did the person foresee the result, was the person coerced to cause the event, etc.

Just as the appraisal of motivational relevance and motivation congruence can be performed as part of the existing Thespian/PsychSim decision-making and belief update processes, we argue here that accountability can be treated as an extension to Thespian/PsychSim's existing approach to support/affinity relationships between agents.

Figure 2 illustrates how an agent can determine accountability for an event's outcome. Our algorithm first looks at the agent whose action directly causes the harm/benefit and judges to what degree it should be held accountable. The algorithm uses the function *IfCoerced*() to determine whether that agent was coerced into performing the given action. If the direct actor was not coerced, it is held fully accountable for the result and the reasoning proceeds no further. Otherwise, any agents that coerced the direct actor are also held partially accountable. The algorithm judges each such coercer on whether it, too, was coerced by somebody else, with any such coercers sharing accountability as well. This process could potentially continue cascading indefinitely, but we instead limit the number of steps for which accountability extends back through the history. In the current investigation, we assume that the observing agent expects others to foresee the effects of their actions. While this assumption is not always true, people do often assume that others will project into the future to the same depth as they will themselves.

Algorithm 2 contains pseudocode for determining whether an agent was coerced, and Algorithm 3 determines who coerced the agent. Currently, we use a qualitative model to judge coercion based on the agent's utility gain/loss. If the agent's chosen action has a strictly greater utility than its other options, then we view the agent's choice as being coerced, in that its decision was driven by its circumstances. If an alternate action would not have decrease the agent's utility, then we view the agent as not being coerced, in that it was free to make a different choice without sacrificing its own utility.

We use a qualitative rather than quantitative method to decide coercion. If all other action options lead to less

Algorithm 2 IfCoerced(*actor*, *pact*)

```
# actor: the agent being studied
# pact: the action performed by actor
# preUtility: actor's utility before doing pact

for action ∈ actor.actionOptions() do
    if action ≠ pact then
        # if there exists another action which does not hurt actor's own utility
        if EU(action) ≥ preUtility then
            return false
return true
```

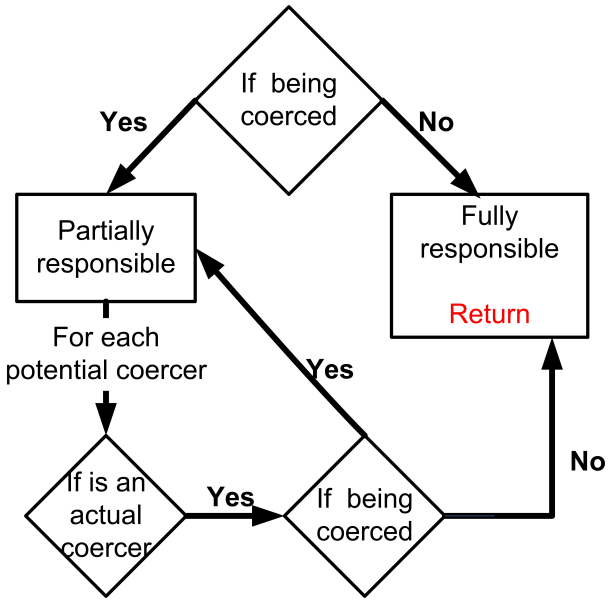

Figure 2: Accountability

utility than its chosen action, then we view the agent as justified in its choice and must look back further to find the accountable agent. In the special case when the agent's chosen action is its best choice but still results in a utility drop, the agent is regarded as not being coerced. In such cases, because the agent is going to be punished regardless of what it does, we treat it as having the freedom to pick actions which will not hurt the other agent's utility.

In Algorithm 2, *preUtility* is defined as in Algorithm 1, with the important distinction that *preUtility* now refers to the beliefs of the observer (the agent who performs appraisal) about the *actor*'s expected utility. Similarly, $EU(action)$ denotes the observer's belief about the *actor*'s utility of alternative option.

To decide who coerced an agent, we consider each agent that acted between the coerced agent's current and last actions. For each potential coercer, if the coerced agent would not have been coerced (as defined by Algorithm 2) if the potential coercer had made a different choice, then the potential coercer is judged as being a true coercer. This

process is illustrated in Algorithm 3.

Algorithm 3 *Is_Coercer_For*(*agent*, *actor*, *agent_pact*, *actor_pact*)

```
# agent_pact: the action performed by agent
# actor_pact: the action performed by actor

for action ∈ agent.actionOptions() do
    if action ≠ agent_pact then
        Simulate action agent_pact
        if not IfCoerced(actor, actor_pact) then
            return true
return false
```

4.3.3 Control

The appraisal of control evaluates the extent to which an event or its outcome can be influenced or controlled by people [30]. It captures not only the individual's own ability to control the situation but also the potential for seeking instrumental social support from other people. Unlike the evaluations of motivational relevance, motivational congruence and accountability in which the most probable mental models of other agents are used for reasoning, here we factor in the probabilities of the mental models because the degree of control is affected by the estimation of how likely certain events will happen in the future.

Algorithm 4 Control(*preUtility*)

```
# preUtility: utility before the event happens

control ← 0
for m1 ∈ mental_models_about_agent1 do
    for m2 ∈ mental_models_about_agent2 do
        for m3 ∈ mental_models_about_self do
            # project limited steps into the future using this set of mental models
            lookahead(m1,m2,m3)
            # curUtility: utility after the lookahead process
            if curUtility ≥ preUtility then
                control ← control + p(m1) * p(m2) * p(m3)
return control
```

Algorithm 4 gives the pseudocode for evaluating control. This algorithm first simulates future steps of the interaction using each possible combination of mental models of self and others, and checks whether the utility drop will be recovered. The algorithm then considers the likelihood of the given mental model combination. For example, assume Granny has two mental models of the wolf. In the first mental model, the wolf will always die after being shot by the hunter. In the second mental model, the wolf will never die even after being shot. Granny believes that there is a 60% possibility that the first mental model is correct. Next assume Granny has two mental models regarding the hunter.

One mental model indicates that the hunter is close by and this mental model has a 50% chance to be correct. The other mental model indicates that the hunter is far away. After Granny is eaten by the wolf, the only event that can help her is the wolf being killed by the hunter. In this case, her sense of control is: $60\% \cdot 50\% = 30\%$.

Algorithm 4 contains pseudocode for a three-agent interaction, but it is straightforward to configure the algorithm for any number of agents. In this algorithm, *preUtility* is defined as in Algorithm 1, but *curUtility* denotes the agent's utility associated with its state after the lookahead projection.

4.3.4 Novelty

In this work, we adapt Leventhal and Scherer's definition of "novelty at the conceptual level", i.e., whether the event is expected from the agent's past beliefs[1] [15, 30]. In our model, novelty appraisal is treated as a byproduct of an agent's belief maintenance. Specifically, in a multiagent context the novelty of an agent's behavior is viewed as the opposite of the agent's motivational consistency, i.e. the more consistent the event is with the agent's motivations, the less novel. Of course, this evaluation is performed from the observing agent's perspective and using the observing agent's beliefs, and there can be discrepancies between what the observing agent feels and what the agent who did the action feels. Computationally, we define novelty as $1 - consistency$, where $consistency$ is calculated as described in Section 2.4.3. The less consistent an action is with the observing agent's expectation about the actor, the higher the novelty if that action happens. For example, if from Red's perspective the wolf did an action which has the second highest utility among the wolf's five alternative actions, the amount of novelty Red will feel if seeing that action is calculated as:

$$1 - \frac{e^3}{\sum_{j=0}^{4} e^j} = 0.37$$

5 Appraisal Model in Action

All the previous examples of our new appraisal model are derived from a Thespian implementation of the Little Red Riding Hood story. In this section we provide two additional scenarios to illustrate the usage of our computational model of appraisal in modeling social interactions. In particular, in Section 5.1 we demonstrate the tight relationship between emotion and cognitive decision-making by showing how appraisal is affected by the depth of reasoning in decision-making. In Section 5.2 we provide a complex situation for accountability reasoning and show that the result of our model is consistent with another validated computational model of social attribution.

[1] Leventhal and Scherer have also defined novelty at the sensory-motor level and schematic level. We did not model them because they are mainly related to people's low level perceptual processes rather than cognitive processes.

Step	Action	Perspective	Lookahead Steps	Motivational Relevance
1	A greets B	B	1	0
		B	2	100
2	B greets A	A	1	0
		A	2	0.99
3	A asks B a question	B	1	0
		B	2	0.99
4	B answers the question	A	1	0
		A	2	0.49

Table 2: Small talk between two persons

5.1 Small Talk

To reveal the tight relationship between cognitive processes and emotion in our model, we implemented an abstract domain of two persons (A and B) taking turns talking to each other. Both of them have these goals: to be talkative and to obey social norms. In fact, just the norm-following behavior itself is an incentive to them — they will be rewarded whenever they do an action that is consistent with social norms. Table 2 contains the two persons' appraisals of motivational relevance regarding each other's actions. We did not include results of other appraisal dimensions as they are less interesting in this scenario.

In PsychSim, we explicitly model the depth of reasoning in agents as the number of steps they project into the future. In this example we provide a comparison of appraisal results when the person's previous reasoning process takes a different number of steps. It can be observed in Table 2 that different depths of reasoning lead to different appraisals. A person appraises another person's initiatives as irrelevant when performing shallow reasoning (lookahead steps = 1). In this case, even though the person has predicted the other person's action, because the action does not bring him/her immediate reward, the person cannot see the relevance of the action. Once the person reasons one step further, he/she finds out that by opening up a topic the other person provides him/her a chance to engage in further conversation and perform a norm following action, the person will then appraise the other person's action as relevant.

5.2 Firing-squad

We implemented the Firing-squad scenario from [17] to illustrate accountability reasoning in which agents are coerced and have only partial responsibility. The scenario goes like this:

In a firing-squad, the commander orders the marksmen to shoot a prisoner. The marksmen refuse the order. The commander insists that the marksmen shoot. They shoot the prisoner and he dies.

We modeled the commander as an agent with an explicit goal of killing the prisoner, and the marksmen as having

Appraisal Dimension	Existing PsychSim Processes
Motivational Relevance	Decision-Making (exploit utility calculations based on single mental model)
Motivational Congruence	Decision-Making (exploit utility calculations based on single mental model)
Accountability	Belief (Mental Model) update (Support Relationship)
Control	Decision-Making (exploit utility calculations based on alternative mental models)
Novelty	Belief (Mental Model) update

Table 3: Appraisal tightly coupled with decision making

no goals related to the prisoner, but they will be punished if they do not obey the commander's order. Using our appraisal model, from the prisoner's perspective, the marksmen hold responsibility for his/her death because they are the persons who directly perform the action. Further, the prisoner can simulate the decision-making process of the marksmen which will lead him/her to find out that the marksmen are coerced because their utilities will be hurt if they do anything else other than shooting. The commander acts right before the marksmen in the scenario and therefore is identified as a potential coercer for the marksmen. Using Algorithm 3 , we can predict that the prisoner can see if the commander chose a different action, the marksmen would not be coerced to shoot. Assuming the prisoner does not find a coercer for the commander, he/she will now believe that the commander holds full responsibility for his/her death. This prediction is consistent with the prediction from Mao's model of social attribution and the data collected from human subjects to validate that model [17].

6 Discussion of Appraisal Model

In PsychSim, comparison among expected utilities plays the central role in decision-making and mental model update. Comparison of expected utilities also plays a central role for deriving appraisal dimensions. Our algorithms for deriving appraisal dimensions demonstrate that no additional calculation of utilities or states other than what has already been performed in the Thespian agent's existing decision-making and belief revision processes is required for appraisal. Table 6 summarizes the relationship between PsychSim's existing processes and the evaluation of appraisal dimensions.

Compared to other computational models of appraisal, the main advantage of this model is that the agents explicitly model other agents' goals, states and beliefs (Theory of Mind). Modeling Theory of Mind makes this model particularly suitable for simulating emotions in social interactions in two ways. First, appraisals are strongly embedded in the social context. For example, novelty is not simply evaluated as whether the physical stimulus is unexpected, but whether the other agents behave as expected. Second, appraisals that are explicitly relevant to social interaction and derivation of social emotion (e.g., accountability) have to leverage Theory of Mind.

Further, the fact that Thespian agents have a Theory of Mind capability enables them to simulate the emotions of others. This ability allows us to simulate an agent's potential mis-expectations about other agents' emotional states. For example, if Granny believes that the hunter can always kill the wolf successfully and the hunter believes that he can only kill the wolf successfully 60% of the time, Granny's control when being eaten by the wolf will be evaluated differently from Granny's and the hunter's perspectives.

The appraisal model can not only simulate ego-centric emotions, but also can simulate emotions that take social relationships into account. Thespian agents can have goals regarding other agents' utilities and emotions (emotion can be modeled as a feature of an agent's state). Therefore, an agent's emotion can be related to other agents' utility changes and emotions. For example, we can simulate an agent having goals of facilitating another agent's goals, or even more specifically having goals of making the other agent feel happy. This agent will act deliberately to help the other agent, and "feel" bad if it hurts the other agent's utility or emotional state.

Finally, the underlying PsychSim framework allows us to explicitly model the depth of reasoning in agents. As shown in Section 5.1, different depths of reasoning lead to different appraisals. Though we have only demonstrated this effect using one appraisal dimension (motivational relevance), this effect is general. Different steps of projection lead to different expectations of future events, and an agent's expectations affect its reasoning about whether an event is novel, whether the effect of the event is changeable and who caused the event.

In terms of future work, one of the extensions we are considering concerns *social appraisal* [16], which argues that another's behaviors, thoughts or feelings can influence an agent's own appraisals. In the current framework, the capacity of the agent to appraise from another's perspectives provides a way for the agent to infer another agent's appraisals. It therefore provides a framework for exploring the various ways those appraisals can influence the agent's own appraisals.

7 Conclusion

We have discussed PsychSim, an environment for multiagent simulation of human social interaction that employs a formal decision-theoretic approach using recursive models. This approach allows us to model phenomena rarely addressed in simulated worlds. We have exploited the recursive models to provide a psychologically motivated computational model of how agents influence each other's beliefs. We have also developed a range of technology to simplify the task of setting up PsychSim models, exploring the simulation and analyzing results. Discussion of these algorithms is beyond the scope of this paper, but they provide users with automated algorithms for fitting simulation parameters to observed behavior, reporting sensitivities in the results, and suggesting potentially interesting perturba-

tions to the scenario [26]. We believe PsychSim has a range of innovative applications, including computational social science and the modeling of social training environments.

As an illustration of such applications, and PsychSim's expressiveness, we provide a computational model of appraisal for POMDP-based agents, implemented in the Thespian framework for interactive narrative. The focus is on five key appraisal dimensions for virtual agents: motivational relevance, motivational congruence, accountability, control and novelty. The approach argues that appraisal is an integral part of a social agent's cognitive processes.

All of these capabilities of the appraisal model derive from the basic PsychSim cognitive components as laid out in Section 2. We were thus able to leverage the decision-theoretic Theory of Mind as implemented in our agents to realize appraisal theory, even though modeling appraisal was not an original intention of the agents' design. The reuse of architectural features therefore provides, not only a novel computational model of emotion, but also a demonstration of a tight relationship between emotion and cognition, suggesting a uniform cognitive structure for emotion and cognition. In addition, by demonstrating how a Theory of Mind capacity is critical to deriving appraisals, and in particular modeling appraisals critical to social emotions like anger, this work argues for the critical role for Theory of Mind in modeling social interaction generally.

8 Acknowledgments

This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM), and the content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

[1] R. P. Abelson, E. Aronson, W. J. McGuire, T. Newcomb, M. Rosenberg, and P. H. Tannenbaum, editors. *Theories of Cognitive Consistency: A Sourcebook*. Rand McNally, Chicago, IL, 1968.

[2] R. Aylett, J. Dias, and A. Paiva. An affectively-driven planner for synthetic characters. In *Proceedings of the International Conference on Automated Planning and Scheduling*, 2006.

[3] R. Cialdini. *Influence: Science and Practice*. Allyn and Bacon, Boston, MA, 2001.

[4] F. de Rosis, C. Castelfranchi, V. Carofiglio, and G. Grassano. Can computers deliberately deceive? a simulation tool and its application to Turing's imitation game. *Computational Intelligence*, 19(3):253–263, 2003.

[5] J. Dias and A. Paiva. Feeling and reasoning: A computational model for emotional characters. In *Proceedings of the Portuguese Conference on Artificial Intelligence*, pages 127–140, 2005.

[6] M. S. El Nasr, J. Yen, and T. Ioerger. Flame: Fuzzy logic adaptive model of emotions. *Autonomous Agents and Multi-Agent Systems*, 3(3):219–257, 2000.

[7] C. Elliott. *The affective reasoner: A process model of emotions in a multi-agent system.* PhD thesis, Northwestern University Institute for the Learning Sciences, 1992.

[8] N. Frijda. *The Emotions.* Cambridge University Press, 1987.

[9] P. J. Gmytrasiewicz and E. H. Durfee. A rigorous, operational formalization of recursive modeling. In *Proceedings of the International Conference on Multi-Agent Systems*, page 125132, 1995.

[10] J. Gratch and S. Marsella. A domain-independent framework for modeling emotion. *Cognitive Systems Research*, 5(4):269–306, 2004.

[11] J. Y. Ito, D. V. Pynadath, and S. C. Marsella. A decision-theoretic approach to evaluating posterior probabilities of mental models. In *AAAI Workshop on Plan, Activity, and Intent Recognition*, 2007.

[12] J. Y. Ito, D. V. Pynadath, L. Sonenberg, and S. C. Marsella. Wishful thinking in effective decision making. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 1527–1528, 2010.

[13] R. S. Lazarus. *Emotion & Adaptation.* Oxford University Press., New York, 1991.

[14] N. Lazzaro. Why we play games: four keys to more emotion in player experiences. In *Game Developers Conference*, 2004.

[15] H. Leventhal and K. R. Scherer. The relationship of emotion and cognition: A functional approach to a semantic controversy. *Cognition and Emotion.*, 1:3–28, 1987.

[16] A. Manstead and A. Fischer. Social appraisal: The social world as object of and influence on appraisal processes. In . T. J. K. R. Scherer, A. Schorr, editor, *Appraisal Processes in Emotion: Theory, Research, Application*, pages 221–232. Oxford University Press, New York, 2001.

[17] W. Mao and J. Gratch. Social causality and responsibility: Modeling and evaluation. In *Proceedings of the International Conference on Virtual Agents*, Kos, Greece, 2005.

[18] R. Marinier, J. Laird, and R. Lewis. A computational unification of cognitive behavior and emotion. *Journal of Cognitive Systems Research*, 2009.

[19] S. C. Marsella and J. Gratch. Ema: A model of emotional dynamics. *Cognitive Systems Research.*, 10(1):70–90, 2009.

[20] S. C. Marsella, D. V. Pynadath, and S. J. Read. PsychSim: Agent-based modeling of social interactions and influence. In *Proceedings of the International Conference on Cognitive Modeling*, pages 243–248, 2004.

[21] R. McAlinden, A. Gordon, H. C. Lane, and D. Pynadath. UrbanSim: A game-based simulation for counterinsurgency and stability-focused operations. In *Proceedings of the AIED Workshop on Intelligent Educational Games*, 2009.

[22] D. Moffat and N. Frijda. Where there's a will there's an agent. In *Proceedings of the ECAI Workshop on Agent Theories, Architectures, and Languages*, Amsterdam, The Netherlands, 1995.

[23] A. Ortony, G. L. Clore, and A. Collins. *The Cognitive Structure of Emotions. Cambridge*. UK: Cambridge University Press, 1998.

[24] A. Paiva, J. Dias, D. Sobral, R. Aylett, P. Sobreperez, S. Woods, and C. Zoll. Caring for agents and agents that care: Building empathic relations with synthetic agents. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*, pages 194–201, 2004.

[25] R. Petty and J. Cacioppo. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer, New York, NY, 1986.

[26] D. V. Pynadath and S. C. Marsella. Fitting and compilation of multiagent models through piecewise linear functions. In *Proceedings of the International Conference on Autonomous Agents and Multi Agent Systems*, pages 1197–1204, 2004.

[27] D. V. Pynadath and S. C. Marsella. PsychSim: Modeling theory of mind with decision-theoretic agents. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1181–1186, 2005.

[28] W. S. Reilly and J. Bates. Building emotional agents. Technical Report CMU-CS-92-143, Carnegie Mellon University, 1992.

[29] I. Roseman. Cognitive determinants of emotion: A structural theory. *Review of Personality and Social Psychology*, 2:11–36, 1984.

[30] K. Scherer. Appraisal considered as a process of multilevel sequential checking. In K. Scherer, A. Schorr, and T. Johnstone, editors, *Appraisal Processes in Emotion: Theory, Methods*. Oxford University Press., Oxford, 2001.

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

[32] K. G. Shaver. *The Attribution Theory of Blame: Causality, Responsibility and Blameworthiness.* Springer-Verlag, 1985.

[33] M. Si, S. C. Marsella, and D. V. Pynadath. Thespian: An architecture for interactive pedagogical drama. In *Proceedings of the Conference on Artificial Intelligence in Education*, 2005.

[34] M. Si, S. C. Marsella, and D. V. Pynadath. Thespian: Using multi-agent fitting to craft interactive drama. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*, pages 21–28, 2005.

[35] M. Si, S. C. Marsella, and D. V. Pynadath. Thespian: Modeling socially normative behavior in a decision-theoretic framework. In *Proceedings of the Conference on Intelligent Virtual Agents*, 2006.

[36] M. Si, S. C. Marsella, and D. V. Pynadath. Proactive authoring for interactive drama: An author's assistant. In *Proceedings of the Conference on Intelligent Virtual Agents*, Paris, France, 2007.

[37] M. Si, S. C. Marsella, and D. V. Pynadath. Directorial control in a decision-theoretic framework for interactive narrative. In *Proceedings of the International Conference on Interactive Digital Storytelling*, pages 221–233, 2009.

[38] M. Si, S. C. Marsella, and D. V. Pynadath. Evaluating directorial control in a character-centric interactive narrative framework. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1289–1296, 2010.

[39] M. Si, S. C. Marsella, and D. V. Pynadath. Modeling appraisal in theory of mind reasoning. *Journal of Autonomous Agents and Multi-Agent Systems*, 20(1):14–31, 2010.

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

[41] C. A. Smith and P. C. Ellsworth. Patterns of appraisal and emotion related to taking an exam. *Personality and Social Psychology*, 52:475–488, 1987.

[42] C. A. Smith and R. S. Lazarus. Emotion and adaptation. In L. A. Pervin, editor, *Handbook of personality:Theory and research.* Guilford, New York, 1990.

[43] J. G. Stacy Marsella and P. Petta. Computational models of emotion. In K. Scherer, T. Bnziger, and E. Roesch, editors, *A blueprint for an affectively competent agent: Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing*. Oxford University Press, Oxford, 2010.

[44] W. Swartout, R. Hill, J. Gratch, W. Johnson, C. Kyriakakis, C. LaBore, R. Lindheim, S. C. Marsella, D. Miraglia, B. Moore, J. Morie, J. Rickel, M. Thiúbaux, L. Tuch, R. Whitney, and J. Douglas. Toward the holodeck: Integrating graphics, sound, character and story. In *Proceedings of the International Conference on Autonomous Agents*, pages 409–416, 2001.

[45] J. Taylor, J. Carletta, and C. Mellish. Requirements for belief models in cooperative dialogue. *User Modelling and User-Adapted Interaction*, 6:23–68, 1996.

[46] D. R. Traum, W. Swartout, S. C. Marsella, and J. Gratch. Fight, flight, or negotiate: Believable strategies for conversing under crisis. In *Proceedings of the Conference on Intelligent Virtual Agents*, 2005.

[47] J. Velasquez. Modeling emotions and other motivations in synthetic agents. In *Proceedings of the National Conference on Artificial Intelligence*, 1997.

[48] B. Weiner. *The Judgment of Responsibility: A Foundation for a Theory of Social Conduct.* The Guilford Press, 1995.

[49] A. Whiten, editor. *Natural Theories of Mind.* Basil Blackwell, Oxford, UK, 1991.