Simultaneously Modeling Humans' Preferences and their Beliefs about Others' Preferences

Sevan G. Ficici
and
Avi Pfeffer

TR-02-07

Computer Science Group
Harvard University
Cambridge, Massachusetts

Simultaneously Modeling Humans' Preferences and their Beliefs about Others' Preferences

Sevan G. Ficici and Avi Pfeffer
Technical Report TR-02-07
Maxwell-Dworkin Laboratory
School of Engineering and Applied Sciences
Harvard University
Cambridge Massachusetts 02138 USA

March 30, 2007 (Revised April 16, 2008)

Abstract

In strategic multi-agent decision making, it is often the case that a strategic reasoner must hold beliefs about other agents and use these beliefs to inform its decision making. The behavior thus produced by the reasoner reflects an interaction between the reasoner's beliefs about other agents and the reasoner's own preferences. In this paper, we are interested to investigate human reasoning, particularly the interaction between a human's utility function and the beliefs the human holds to reason about other agents. A significant challenge faced by model designers, therefore, is how to model such a reasoner's behavior so that the reasoner's preferences and beliefs can each be identified and distinguished from each other. In this paper, we introduce a model of strategic human reasoning that allows us to distinguish between the human's utility function and the human's beliefs about another agent's utility function as well as the human's beliefs about how that agent might interact with yet other agents. We show that our model is uniquely identifiable. We then illustrate the performance of our model in a multi-agent negotiation game.

1 Introduction

Many multi-agent domains involve both human and computer decision makers that are engaged in collaborative or competitive activities. Examples include online auctions, financial trading, scheduling, and computer gaming (online and video). To construct computer agents that can interact successfully with human participants, we need to understand several things about human reasoning in multi-agent domains. Behavioral economics [Rabin, 1998, Camerer, 2003] has shown that people employ social utility functions that deviate from rational game-theoretic prescription. Learning about the social utilities humans use has been shown to be beneficial for the design of computer agents [Gal et al., 2004]. Psychological theories of mind explore people's reasoning about others [Gordon, 1986, Davies and Stone, 1995, Hurley, 2004, Cacioppo et al., 2005]; one aspect of a theory of mind concerns *beliefs* about others. This aspect of modeling human reasoning for use in computer agents is left open by [Gal et al., 2004].

We are interested to investigate beliefs that human reasoners hold about other agents. Do people hold beliefs about another agent's preferences or intentions, and use these beliefs to inform decision making? If so, then what are these beliefs? Are these beliefs correct? Do the beliefs follow a pattern of some sort? How do the beliefs about others' preferences or intentions relate to the preferences or intentions of the reasoner? Do people believe others to be the same as themselves? If people use beliefs to reason, then their behavior is the result of an interaction between their beliefs about others and their own utility function; can we distinguish between the two and untangle a person's beliefs from her utility function? For example, the negotiation experiments of [Gal et al., 2004] indicate that human players often make offers that are more generous than necessary to be accepted. This result may indicate a gap between the proposer's beliefs about the responder and the responder's actual behavior. Alternatively, the human player may have a strong preference to be generous. If we are interested to identify an explanation, we must construct a model such that we can distinguish the reasoner's preferences from its beliefs about another agent. Otherwise, model parameters will represent some amalgam of these various factors.

In this paper, we use AI techniques to address the above questions. We introduce a model of strategic human reasoning that allows us to distinguish between three factors: the human's own utility function, the human's beliefs about another agent's utility function, and the human's beliefs about how that other agent may interact with yet other agents. To support our claim that the model allows us to untangle these factors, we show that the model is uniquely identifiable; that is, no two different parameters sets can produce the same model behavior over all possible inputs. We provide a learning algorithm for our model and analyze how well learned models fit data obtained from human-subjects trials of a multi-agent negotiation game. In addition to investigating our general model, we also examine constrained versions that correspond to particular belief patterns. For example, it may be that a human player believes other agents to share the same utility function as the human. In another example, we consider the case where different humans have different personal preferences, but share the same beliefs about the preferences of other agents. Our analyses provide insight into whether modeling a person's beliefs about others' preferences separately from the person's own preferences yields a better model for use in computer agents.

Our work is significantly different from most work in multi-agent systems (MAS) [Weiss, 2000, Kraus, 2001]. For example, MAS research often focuses on environments comprised of only computer agents; thus, agents tend to be viewed as rational actors [Gmytrasiewicz and Durfee, 2000]. The boundedrational agents of [Vidal and Durfee, 1995] do not specifically address human boundedness. Models of human emotion [Gratch and Marsella, 2005] are generally not learned from real human data. Finally, investigations of theories of mind do not address the construction of computer agents that are to interact with humans.

2 Negotiation Game

To investigate the role of beliefs about others in human reasoning, we require certain elements in a multi-agent environment. First, we require a domain where agents' preferences matter for decision making. The domain should provide the possibility for agents to reason strategically about each other, and this reasoning may entail the need for beliefs. Agents in the domain should be *situated* [Lueg and Pfeifer, 1997], such that behavior requires interaction within and with the environment; [Allain, 2006] shows that situated task activity elicits stronger concern with social factors such as fairness, whereas the same underlying game presented in a more abstract payoff-matrix form engenders behavior more in line with rational Nash-equilibrium play. Finally, we require the domain to be simple enough for modeling and decision making to be tractable. The Colored Trails (CT) framework [Grosz et al., 2004] meets our requirements.

Using the CT framework, we construct a negotiation game in which players must negotiate with each other to obtain resources needed to complete a task. The player we are interested to model is called a *proposer*; a proposer formulates an offer to exchange resources with another player who is called the *responder*. The responder may also receive an outside offer to exchange resources. The responder can accept only one offer, either the proposer's or some outside offer, or the responder can reject all received offers. Our domain can be viewed as a general model for one-shot negotiation that is situated in a particular task.

3 Player Models

We are interested to model the behavior of proposer agents in our game. In particular, we model the proposer as maintaining beliefs about how the responder will behave and using these beliefs to reason about what offer to make. Thus, our proposer model contains parameters that facilitate reasoning about its own preferences as well as other parameters that facilitate reasoning with beliefs about the responder. Specifically, the proposer uses its beliefs about the responder to calculate the expected utilities of the possible offers it can make. In this section, we introduce our general model for representing a proposer's beliefs about the responder.

Our models make use of only two simple features that quantify proposal properties; these features are rather general and can be applied to almost any negotiation game. Let *self-benefit* (SB) quantify the change in score a player will receive if a proposal is accepted, and *other-benefit* (OB) quantify the change in score the other player will receive if the proposal is accepted.

Let each proposal $O = \langle \mathrm{SB}, \mathrm{OB} \rangle$ be a vector of feature values; let $\mathbf{w} = \langle w_{\mathrm{SB}}, w_{\mathrm{OB}} \rangle$ and $\mathbf{v} = \langle v_{\mathrm{SB}}, v_{\mathrm{OB}} \rangle$ be vectors of feature weight parameters. The parameters in $\mathbf{w}$ are those which the proposer uses to reason about its own preferences; the parameters in $\mathbf{v}$ are those which the proposer uses to reason about the responder's believed preferences. Preference is expressed by a utility function $U : \mathbb{O} \rightarrow \mathbb{R}$ on the space $\mathbb{O}$ of offers, which computes a linear combination of feature values using a set of weights.

Let ϕ denote the *status-quo*, which for the responder represents the option of rejecting the proposer's offer as well as the outside offer, and for the proposer represents the proposal that no resources change hands. Note that $U(\phi) = 0$, since $\mathrm{SB} = \mathrm{OB} = 0$.

Since not all humans will likely share the same preferences and beliefs about others' preferences, we use mixture models to cluster human play into different behavioral types. Let ρ^{s_i} be the proportion of proposers of type s_i. A proposer of type s_i uses the utility function U^{s_i} with weight vector $\mathbf{w}^{s_i}$ to reason about its own preferences:

$$U^{s_i}(O) = \sum_l w_l^{s_i} \cdot O_l. \tag{1}$$

A proposer of type s_i may believe that different types of responders exist. Let ρ^{s_i,t_j} be the proportion of responders of type $\{s_i, t_j\}$ that a proposer of type s_i believes to exist. A proposer of type s_i uses the utility function U^{s_i,t_j} with weight vector $\mathbf{v}^{s_i,t_j}$ to reason about the preferences it believes a responder of type $\{s_i, t_j\}$ has:

$$U^{s_i,t_j}(O) = \sum_l v_l^{s_i,t_j} \cdot O_l. \tag{2}$$

This is different from other work [Gal et al., 2004, Gal and Pfeffer, 2006] because the type $\{s_i, t_j\}$ of the responder is embedded in the type s_i of the proposer, who is holding a belief about the preferences of the responder.

Humans select offers non-deterministically; we are prone to make errors. Further, since our models will not be perfect, we care to have our models attach probabilities to different outcomes. To accommodate these factors, we convert proposal utilities to probabilities of selection with a multinomial logit function. The probability that a responder of type $\{s_i, t_j\}$ accepts an offer O is

$$\overset{\text{Responder}}{\Pr}(\text{accept}|O, \phi, \{s_i, t_j\}) = \frac{e^{U^{s_i,t_j}(O)}}{e^{U^{s_i,t_j}(O)} + e^{U^{s_i,t_j}(\phi)} + e^{z}}. \tag{3}$$

Since we do not know the outside offer, we cannot compare with the utility of a specific offer; rather, we use a parameter z to represent the believed utility of a generic, unknown outside offer. Taking an expectation over all responder types t_j that a proposer of type s_i believes to exist gives

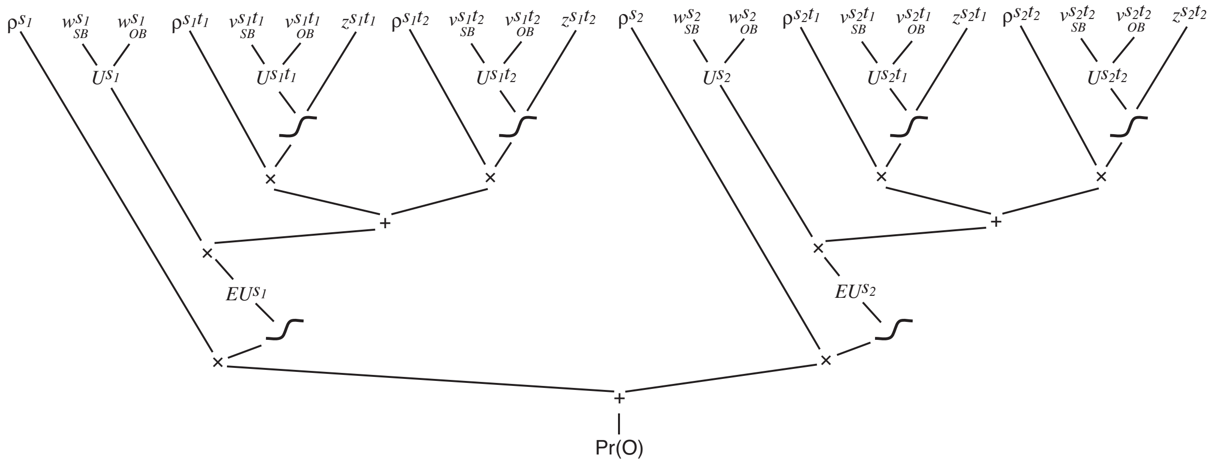

Figure 1: Structure of general model. Two proposer types, each with two responder types embedded within, are shown. Model parameters are located at the leaves of the diagram.

$$\overset{\text{Responder}}{\Pr}(\text{accept}|O,\phi,s_i)= \sum_j \overset{\text{Responder}}{\Pr}(\text{accept}|O,\phi,\{s_i,t_j\}) \cdot\rho^{s_i,t_j}. \tag{4}$$

For the proposer, let $\mathcal{O} = \{O_1, \ldots, O_M\}$ be the set of possible offers, where M varies from game to game due to the particulars of the game state. The probability that a proposer of type s_i will select the m-th proposal in $\mathcal{O}$ is a function of the expected utility of O^m:

$$EU^{s_i}(O^m) = U^{s_i}(O^m)\cdot \overset{\text{Responder}}{\Pr}(\text{accept}|O^m,\phi)\ . \tag{5}$$

The proposer obtains the expected utility of an offer O by weighting the utility of O, if accepted, to the proposer by the believed probability that the responder will accept O. With expected utilities in hand, the probability that a proposer of type s_i will select the m-th proposal in $\mathcal{O}$ is

$$\overset{\text{Proposer}}{\Pr}(\text{selected} = O^m|\mathcal{O},s_i)= \frac{e^{EU^{s_i}(O^m)}}{\sum_k e^{EU^{s_i}(O^k)}}. \tag{6}$$

Taking an expectation over proposer types gives

$$\overset{\text{Proposer}}{\Pr}(\text{selected} = O^m|\mathcal{O})= \sum_i \overset{\text{Proposer}}{\Pr}(\text{selected} = O^m|\mathcal{O},s_i) \cdot\rho^{s_i} \tag{7}$$

The general model's structure, for two proposer types and two embedded responder types in each proposer type, is illustrated in Figure 1.

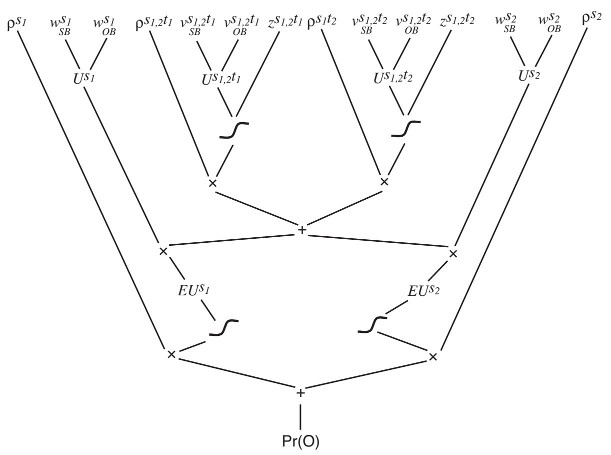

Figure 2: Structure of constrained model where different proposer types share beliefs about the responder types.

4 Hypotheses and Model Variations

We explore four hypotheses with our model. The first hypothesis is that proposer beliefs about responders are correct, and match the actual observed behavior of human responders. To explore this possibility, we pre-learn a model of responder behavior from responder data; we regard this responder model to have the correct values for parameters v_l and z, as well as the correct mixture distribution over responder types (in building the responder model, we determined that two mixture components works best). We then fix these parameters in our general model and learn only the proposer parameters $w_l^{s_i}$ and ρ^{s_i}.

The second hypothesis is that proposer beliefs are incorrect. To explore this possibility, we learn all of the parameters of our general model simultaneously. If we obtain a better fit of proposer data under this approach than above, then proposer beliefs are incorrect.

With respect to the *pattern* of beliefs, our third hypothesis is that all proposer types believe the same thing about responders. Here the different proposer types may have different utility functions to express their own preferences, but they have the same beliefs about the responders. Thus, $v^{s_0,t_j} = v^{s_1,t_j} = \ldots = v^{s_S,t_j}$, and similarly for parameters z^{s_i,t_j} and ρ^{s_i,t_j}. This structure is illustrated by Figure 2.

Our fourth hypothesis is that people believe other people are like themselves. Specifically, each proposer type believes that the responder has the same utility function as itself. In this case, each proposer type s_i believes that there exists one responder type $\{s_i, t_1\}$ and that $v_{\mathrm{SB}}^{s_i,t_1} = w_{\mathrm{SB}}^{s_i}$ and $v_{\mathrm{OB}}^{s_i,t_1} = w_{\mathrm{OB}}^{s_i}$. This structure is illustrated by Figure 3.

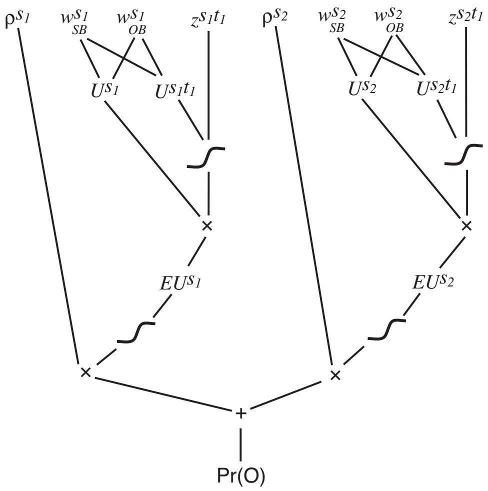

Figure 3: Structure of constrained model where each proposer type believes the responder type has the same preferences as itself.

5 Identifiability

If a proposer makes a generous offer to the responder, we are left to wonder whether the proposer believes the responder will accept nothing less, or whether the proposer has a preference to be generous. Both the proposer's preferences and beliefs about the responder have an effect on proposer behavior. Can we untangle the effects of one from the other? In other words, are the subjective beliefs proposers have about responders identifiable? It is not obvious that they are. Nevertheless, in this section, we show that, at least theoretically, preferences can be distinguished from beliefs in each proposer type s^i. We are only interested to show this within one proposer type. It may be possible that different proposer mixtures produce the same behavior; nevertheless, in all models, the effects of the proposer's preferences, her beliefs about the responder's preferences, and her beliefs about the outside offer can be distinguished. No more than one set of parameter values can produce the same model behavior over all possible inputs.

To begin, let us use a minimal model with one responder type, and one generic proposal feature O_x and associated parameters w, v, and z. Our proof requires only that we look at the proposer's expected utility (5), which is

$$\begin{aligned} EU(O) &= (w \cdot O_x) \cdot \frac{e^{v \cdot O_x}}{e^{v \cdot O_x} + e^{v \cdot \phi_x} + e^z} \\ &= (w \cdot O_x) \cdot \frac{1}{1 + e^{-v \cdot O_x} + e^{z - v \cdot O_x}}. \end{aligned} \tag{8}$$

Let us first consider parameters values $w > 0$, $v > 0$. Let us assume that two parameter sets $\langle w_0, v_0, z_0 \rangle$ and $\langle w_1, v_1, z_1 \rangle$, where $w_0 \neq w_1$, $v_0 \neq v_1$, or

$z_0 \neq z_1$, produce the same value in (8) for all feature values O_x. Let h denote the probability that the responder will accept an offer (3). Since (3) is a sigmoid function, $\lim_{O_x \to \infty} h(accept|O, \phi) \to 1.0$. Thus, for sufficiently large values of O_x, Equation (8) approximates a linear equation with slope w. Clearly, two different values of w will create different lines with different slopes for large O_x. Thus, two parameter sets with different values for w cannot produce identical behavior in (8) over all O_x; as a result, it must be that $w_0 = w_1$. This reasoning holds even when we have multiple responder types: if the responder types have different but similar preferences, we merely locate an offer that takes all of their sigmoid functions to 1.0; if the responder types have opposing preferences, then we locate an offer that takes some sigmoids to 1.0 and others to 0.0, at which point the mixture outputs a constant between 0.0 and 1.0. In either case, the expected utility function becomes a line.

Next, we examine the denominator of the exponent in (8). Parameters v and z interact in the denominator, but $z - v \cdot O_x$ is a line. No two pairs of $\langle v_0, z_0 \rangle$ and $\langle v_1, z_1 \rangle$ can describe the same line, unless $v_0 = v_1$ and $z_0 = z_1$; thus, the effects of parameters v and z can be distinguished from each other.

Because h is a sigmoid function, the parameters v and z together determine the width of the domain interval over which h spans the range interval $[0 + \epsilon, 1 - \epsilon]$. Given different v and z, we will obtain different transitions in the sigmoid. Therefore, it must also be that $v_0 = v_1$ and $z_0 = z_1$. This completes our proof for $w, v > 0$. For parameter values where one or both of w or v is less than zero, our reasoning is identical. When $v < 0$, we instead have $\lim_{O_x \to -\infty} h(accept|O, \phi) \to 1.0$. When $w < 0$, we instead have a negative slope.

Now let us assume that proposals have multiple features. For any two vectors $\mathbf{w}_0$ and $\mathbf{w}_1$, the proposer's utility function U can only produce the same values over all possible offers if the corresponding elements of $\mathbf{w}_0$ and $\mathbf{w}_1$ are identical. If $\mathbf{w}_0$ and $\mathbf{w}_1$ are not identical, then we can construct an offer with value zero at each feature except for where the corresponding weights in $\mathbf{w}_0$ and $\mathbf{w}_1$ are different, and we will obtain different utilities. Identical reasoning applies to the responder feature vector $\mathbf{v}$.

6 Learning

Models are trained by gradient descent. Let $g(\text{selected} = O^* | \mathcal{O})$ be the probability that some proposer model assigns to the proposal O^* that was actually selected by a human proposer, given the set of options $\mathcal{O}$. The error function F that we minimize measures negative log likelihood of the data (N instances), given a model:

$$F = -\sum_{n=1}^{N} \ln \Big(g(\text{selected} = O^{*n} | \mathcal{O}^n) \Big). \tag{9}$$

The derivative of the error function with respect to some model parameter

$w_l^{s_i}$ (or $v_l^{s_i,t_j}$, z^{s_i,t_j}, or ρ^{s_i,t_j}) is

$$\frac{\partial F}{\partial w_l^{s_i}} = -\sum_{n=1}^{N} \frac{\frac{\partial g}{\partial w_l^{s_i}}}{g(\text{selected} = O^{*n}|\mathcal{O}^n)}. \qquad (10)$$

Equation (10) requires that we further calculate the partial derivative of function g. The partial derivative of g with respect to the proposer's feature weights w^{s_i} is

$$\frac{\partial g(\text{selected} = O^*|\mathcal{O})}{\partial w_l^{s_i}} =$$

$$g(\text{selected} = O^*|\mathcal{O}, s_i) \cdot \Big(W(O^*) - \sum_k W(O^k) \cdot g(\text{selected} = O^k|\mathcal{O}, s_i)\Big), \qquad (11)$$

where

$$W(O) \equiv O_l \cdot \overset{\text{Responder}}{\Pr}(\text{accept}|O, \phi, s_i) \qquad (12)$$

The partial derivative of g with respect to the proposer's beliefs about the responder's feature weights v^{s_i,t_j} is

$$\frac{\partial g(\text{selected} = O^*|\mathcal{O})}{\partial v_l^{s_i,t_j}} =$$

$$g(\text{selected} = O^*|\mathcal{O}, s_i) \cdot \Big(Q(O^*) - \sum_k Q(O^k) \cdot g(\text{selected} = O^k|\mathcal{O}, s_i)\Big) \qquad (13)$$

where

$$Q(O) \equiv$$

$$U^{s_i}(O) \cdot \overset{\text{Responder}}{\Pr}(\text{accept}|O, \phi, \{s_i, t_j\}) \cdot \rho^{s_i,t_j} \cdot \Big(O_l - O_l \cdot \overset{\text{Responder}}{\Pr}(\text{accept}|O, \phi, \{s_i, t_j\})\Big) \qquad (14)$$

To calculate gradients for the embedded responder parameters, we compute $Q(O)$. The first line on the right-hand side of (14) is the contribution a responder of type $\{s_i, t_j\}$ makes to the expected utility of a proposer of type s_i; the second line is an adjustment factor equal to the difference, given the responder type, between the feature value O_l if accepted and its expected value.

The partial derivative of g with respect to the proposer's belief about the responder's parameter z, which represents the utility of a generic outside offer $\overline{O}$, is:

$$\frac{\partial g(\text{selected} = O^*|\mathcal{O})}{\partial z_l^{s_i,t_j}} = g(\text{selected} = O^*|\mathcal{O}, s_i)\cdot \Big(R(O^*) - \sum_k R(O^k) \cdot g(\text{selected} = O^k|\mathcal{O}, s_i)\Big), \tag{15}$$

where

$$R(O) \equiv U^{s_i}(O) \cdot \rho^{s_i,t_j} \cdot \overset{\text{Responder}}{\Pr}(\text{accept } \overline{O}|O, \phi, \{s_i, t_j\}) \cdot \overset{\text{Responder}}{\Pr}(\text{accept } \overline{O}|O, \phi, \{s_i, t_j\}) \tag{16}$$

The partial derivative of g with respect to the proposer's beliefs about the distribution over responder types is

$$\frac{\partial g(\text{selected} = O^*|\mathcal{O})}{\partial \rho^{s_i,t_j}} = g(\text{selected} = O^*|\mathcal{O}, s_i)\cdot \Big(Z(O^*) - \sum_k Z(O^k) \cdot g(\text{selected} = O^k|\mathcal{O}, s_i)\Big) \tag{17}$$

where

$$Z(O) \equiv U^{s_i}(O)\cdot \overset{\text{Responder}}{\Pr}(\text{accept}|O, \phi, \{s_i, t_j\}) \tag{18}$$

Let α be our learning rate. The weight-update equation for some model parameter $w_l^{s_i}$ (or, $v_l^{s_i,t_j}$, z^{s_i,t_j}) is

$$w_l^{s_i} \leftarrow w_l^{s_i} - \alpha \cdot \frac{\partial F}{\partial w_l^{s_i}}. \tag{19}$$

To update the probability ρ^{s_i} of proposer type s_i in the mixture, we multiply by the gradient, which turns out to be equivalent to using Bayes' rule:

$$\begin{aligned} \rho^{s_i} &\leftarrow -\frac{\partial F}{\partial \rho^{s_i}} \cdot \rho^{s_i} \\ &= \sum_{n=1}^{N} \frac{g(\text{selected} = O^{*n}|\mathcal{O}^n, s_i) \cdot \rho^{s_i}}{g(\text{selected} = O^{*n}|\mathcal{O}^n)}. \end{aligned} \tag{20}$$

Finally, to update the beliefs a proposer of type s_i has about the distribution over responder types $\{s_i, t_j\}$, we use the following:

$$\rho^{s_i,t_j} \leftarrow \rho^{s_i,t_j} - \alpha \cdot \frac{\partial F}{\partial \rho^{s_i,t_j}} \tag{21}$$

The equations given here assume the general model. To learn one of the model variations above where parameter weights are shared, we combine the gradients for all the parameters involved in an equivalence and use the combined gradient to update all the equivalent parameters.

6.1 Results

We ran human-subjects trials to collect data on how humans play our negotiation game; we collected 536 data instances. We used cross-validation to determine how well the model variations described above fit human proposer data; to speed training, we initialized all models with weights from our model of Hypothesis 1, which was trained first on responder data then proposer data. Results are summarized in Table 1, where each column gives the average negative log likelihood (NLL) of the data, given the model, divided by the size of the test data; lower numbers indicate better fit. Our learning process sought to minimize this measurement. The best fit is obtained under the model constructed to test hypothesis H2, which asks whether proposers' beliefs about the responders are incorrect. Since Hypothesis 2 is investigated by using the general model, it should fit our data no worse than any other model, and this is indeed the case.

Table 1: Fit of learned models to data test-sets.

	1	2	3	4	5	6
	Random	H4	Reflexive	H1	H3	H2
NLL	5.0106	4.4991	4.4220	4.0873	4.0803	**4.0698**

Next, we have hypotheses H3 and H1, which ask whether proposers have shared beliefs about responders, and whether proposers have correct beliefs about responders, respectively. The model used to investigate Hypothesis 3 (Figure 2) is a generalization of the model used to investigate H1 (correct beliefs imply shared beliefs). Thus, the model in Figure 2 should fit our data no worse than the model in H1, which is also what our data show.

The difference in fit between H2 and H1 is small but has a p-value of 0.0707 (single-tailed t-test). Thus, there is a suggestion that proposers have slightly incorrect beliefs about responders. The p-value when comparing the fit of H2 and H3 is 0.1418. This is weaker evidence that proposers do not share the same beliefs.

Hypothesis H4 asks whether proposers believe that responders share the same preferences as themselves. This hypothesis clearly has the worst fit of our hypotheses and so appears false. Column 3 of Table 1 gives the performance of a reflexive model of proposer behavior; this model makes a decision based only upon the proposal options it has, and does not explicitly reason about the responder at all. The reflexive proposer model is a mixture model identical to our general model, except that it contains none of the parameters that relate to the responder. Interestingly, the model of hypothesis H4 fits our data worse than the reflexive model. That is, having a model with poor beliefs can be worse than a model with no beliefs at all.

Finally, column 1 of Table 1 gives the performance we may expect from a model that uses a uniform distribution over all proposals in $\mathcal{O}$, and so corresponds to random guessing. All of our models clearly fit our data better than random guessing.

6.2 Conclusions

Human decision making in multi-agent scenarios is the product of several factors, such as individual preference, beliefs about others' preferences, and beliefs about how others interact with third parties. We have introduced the first model of human reasoning that allows us to identify and distinguish these factors; we show that our model is identifiable. Using a simple multi-player negotiation game, we conduct human-subjects trials to obtain data about human reasoning. We then use cross validation to determine how well each of several variations of our general model fit our data; each model variation corresponds to a particular hypothesis we investigate. These hypotheses ask whether people form correct beliefs of others' preferences or not, and whether one's beliefs relate to one's preferences in particular ways. We find that, in our negotiation game, players form slightly incorrect beliefs about each other's preferences; current work investigates whether we can sharpen these results with improved learning. Our results have implications for agent designers who want to build computer agents that interact with people in strategic situations. Examining human beliefs in other domains will be helpful future work.

Acknowledgments

The research reported in this paper was supported in part by NSF grant CNS-0453923 and AFOSR grant FA9550-05-1-0321. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, AFOSR, or the U.S. Government.

References

[Allain, 2006] Allain, F. A. (2006). The effect of context on decision making in Colored Trails. Technical report, Harvard College. Undergraduate honors thesis.

[Cacioppo et al., 2005] Cacioppo, J. T., Visser, P. S., and Pickett, C. L., editors (2005). *Social neuroscience: People thinking about people.* MIT Press.

[Camerer, 2003] Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction.* The Roundtable Series in Behavioral Economics. Princeton University Press.

[Davies and Stone, 1995] Davies, M. and Stone, T., editors (1995). *Folk Psychology: The Theory of Mind Debate.* Blackwell Publishers.

[Gal and Pfeffer, 2006] Gal, Y. and Pfeffer, A. (2006). Predicting people's bidding behavior in negotiation. In *Autonomous Agents and Multi-Agent Systems.*

[Gal et al., 2004] Gal, Y., Pfeffer, A., Marzo, F., and Grosz, B. J. (2004). Learning social preferences in games. In *National Conference on Artificial Intelligence (AAAI).*

[Gmytrasiewicz and Durfee, 2000] Gmytrasiewicz, P. and Durfee, E. H. (2000). Rational coordination in multi-agent environments. *Autonomous Agents and Multi-Agent Systems Journal*, 3(4):319–350.

[Gordon, 1986] Gordon, R. (1986). Folk psychology as simulation. *Mind and Language*, 1:158–171.

[Gratch and Marsella, 2005] Gratch, J. and Marsella, S. (2005). Evaluating a computational model of emotion. *Autonomous Agents and Multi-Agent Systems*, 11(1):23–43.

[Grosz et al., 2004] Grosz, B. J., Kraus, S., Talman, S., Stossel, B., and Havlin, M. (2004). The influence of social dependencies on decision-making: Initial investigations with a new game. In *Autonomous Agents and Multi-Agent Systems.*

[Hurley, 2004] Hurley, S. (2004). The shared circuits hypothesis: A unified functional architecture for control, imitation, and simulation. In Hurley, S. and Chater, N., editors, *Perspectives on Imitation: From Neuroscience to Social Science*, volume 1. MIT Press.

[Kraus, 2001] Kraus, S. (2001). *Strategic Negotiation in Multiagent Environments.* MIT Press.

[Lueg and Pfeifer, 1997] Lueg, C. and Pfeifer, R. (1997). Cognition, situatedness, and situated design. In *Conference on Cognitive Technology.*

[Rabin, 1998] Rabin, M. (1998). Psychology and economics. *Journal of Economic Literature*, 36:11–46.

[Vidal and Durfee, 1995] Vidal, J. M. and Durfee, E. H. (1995). Recursive agent modeling using limited rationality. In *International Conference on Multi-Agent Systems.*

[Weiss, 2000] Weiss, G., editor (2000). *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. MIT Press.