The Interplay of Emotions and Norms in Multiagent Systems

Anup K. Kalia[1] , **Nirav Ajmeri**[2] , **Kevin S. Chan**[3] , **Jin-Hee Cho**[4] ,
Sibel Adalı[5] and **Munindar P. Singh**[2]
[1]IBM Thomas J. Watson Research Center
[2]North Carolina State University
[3]US Army Research Lab
[4]Virginia Polytechnic Institute and State University
[5]Rensselaer Polytechnic Institute
anup.kalia@ibm.com, najmeri@ncsu.edu, kevin.s.chan.civ@mail.mil, jicho@vt.edu, sibel@cs.rpi.edu, mpsingh@ncsu.edu

Abstract

We study how emotions influence norm outcomes in decision-making contexts. Following the literature, we provide baseline Dynamic Bayesian models to capture an agent's two perspectives on a directed norm. Unlike the literature, these models are holistic in that they incorporate not only norm outcomes and emotions but also trust and goals. We obtain data from an empirical study involving game play with respect to the above variables. We provide a step-wise process to discover two new Dynamic Bayesian models based on maximizing log-likelihood scores with respect to the data. We compare the new models with the baseline models to discover new insights into the relevant relationships. Our empirically supported models are thus holistic and characterize how emotions influence norm outcomes better than previous approaches.

1 Introduction

Agents interact with each other to make informative decisions. An agent's emotions can be influenced by whether its goals are achieved and whether norms are satisfied by other agents. Consider two agents, Alice and Bob. Suppose Alice has a goal to complete a task that she cannot complete by herself. She requests Bob to complete the task. Bob agrees, meaning he creates a commitment norm toward Alice to perform that task. Now consider two possibilities.

Example 1 *Bob completes the task and, thus, satisfies his commitment toward Alice. The satisfaction of the commitment leads Alice to achieve her goals. Alice is now happy, her trust for Bob increases, and she may display positive emotions toward Bob. Bob, on receiving Alice's positive feedback, may feel encouraged to interact again with her. And, Alice may as well—possibly commit to Bob for something else.*

Example 2 *Bob fails to complete the task he committed to perform, thereby violating his commitment to Alice. Alice might be unhappy since she fails to achieve her goals. If she blames Bob, she may lose trust in Bob. And, each of them may be disinclined to commit to the other in the future.*

Examples 1 and 2 suggest that Alice's decision to commit to Bob depends on the outcomes of her prior interactions with him, including how her goals turned out, her appraisal of the situation (e.g., blame Bob?), and her resulting emotions.

Works on norm recommendation are geared to norm emergence [Brooks *et al.*, 2011; Mahmoud *et al.*, 2016; Ajmeri *et al.*, 2018] based on prior outcomes of norms, including sanctions such as rewards or punishments. But, real-life sanctions are more subtle, including change of trust or emotions that might influence an agent's actions [Nardin *et al.*, 2016]. Thus, it is important to consider norm outcomes with respect to emotions, trust, and goals. Existing works capture relationships among these variables: emotions and trust [Dunn and Schweitzer, 2005; Antos *et al.*, 2011; Paradeda *et al.*, 2017], trust and commitments [Kalia *et al.*, 2014], emotions and goals [Guiraud *et al.*, 2011; Lallé *et al.*, 2018], and goals and commitments [Telang *et al.*, 2019]. However, they lack a holistic view of the relationships.

Developing a suitable holistic model is nontrivial. To make the problem tractable, one, we limit our scope to commitments as a type of norm. Two, we simplify our treatment of emotions to reduce the complexity of our study design. Specifically, we adopt a simple form of appraisal theory [Arnold, 1960; Lazarus, 1966] assuming that agents process their emotions discretely (not continuously) by appraising the state changes of their goals and norms. Three, we adopt a simple form of Dimensional Theory [Russell, 1980] and of the Ortony, Clore, & Collins (OCC) model of emotions [Ortony *et al.*, 1988], in which we consider valence for stating emotions but not arousal or intensity.

Contributions. We propose a novel empirical analysis of emotions and norms based on an extension of the well-known *Colored Trails* [Gal *et al.*, 2010]. We map aspects of this game to variables such as norms, emotions, trust, and goals. We use the data we collect from the game play to generate Dynamic Bayesian models in a step-wise process to capture relationships between these variables. We evaluate these relationships via log-likelihood scores with respect to the data to understand whether these relationships hold. Further, for each target variable, we identify the relationship that yields the highest prediction accuracy.

2 Related Work

We now describe the key related work. From virtual agents, De Melo et al. [2012] provide a model that considers the interpersonal effect of emotion in decision-making. Antos et al. [2011] and Paradeda et al. [2017] endorse the importance of emotions in trust. Hoegan et al. [2017] show how emotions influence social decisions. Sébastien et al. [2018] provide evidence that a student's emotions are modulated by the student's achievement goals. Such contributions focus on decision-making and do not formalize their outcomes in terms of norm or goal satisfaction. The lack of formalization makes it difficult to enhance a model with additional dependent variables.

In terms of multiagent systems, Dastani and Lorini [2012] associate emotions with goals. Steunebrink et al. [2007] formalize emotions in terms of agents' beliefs, goals, abilities, plans, intentions, and commitments. Lorini and Schwarzentruber [2009] capture emotions as the difference between the outcomes of current choices made and choices that could have been made. Existing formal models represent only a portion of the relationships. Importantly, these approaches neither provide nor evaluate computational models for predicting norm outcomes. Kalia et al.'s [2014] is another multiagent approach. They provide some experimental evaluations, but their model is limited to trust and commitments.

In the area of norm emergence, Brooks et al. [2011] capture how agents learn from their own experiences to converge to a behavior that becomes a norm in their society. Mahmoud et al. [2016] describe how punishment by peers leads to norm emergence. Nardin et al. [2016] suggest self-directed sanctions (e.g., guilt and trust) and other-directed sanctions (e.g., gossip and praise). Ajmeri et al. [2018] develop personal agents that infer contextually relevant norms on observing norm deviations and understanding the social context related to the deviation and applied sanctions. In psychology, Dunn and Schweitzer [2005] describe the influence of emotional states on trust. Forgas [1995] proposes how a human's emotions influence his or her judgments.

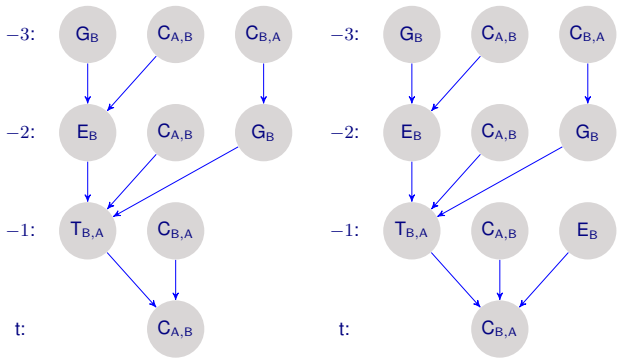
Overall, the above models do not provide a holistic picture of relationships between norms, goals, emotions, and trust.

3 Conceptual Framework

We treat commitments, goals, trust, and emotions as four discrete random variables in our Bayesian model.

Commitment $\mathsf{C}_{\mathsf{A,B}}$. A commitment $\mathsf{C}_{\mathsf{A,B}}$ means that a debtor A commits to a creditor B to bring about a consequent provided an antecedent holds. A commitment provides grounds for B to expect some actions from A [Singh, 1999]. The outcome of a commitment can be represented as: *satisfied* (sat) when the consequent holds regardless of whether the antecedent does; or *violated* (vio) when the antecedent holds but the consequent fails to hold.

Goal G_{A}. A goal is a condition that an agent wants to achieve and may motivate the agent to act, but is not directly visible to others. The outcome of a goal G_{A} has a binary value, *achieved* (ach) or *failed* (fai).



(a) Model M_1 for $\mathsf{C}_{\mathsf{A,B}}$. (b) Model M_2 for $\mathsf{C}_{\mathsf{B,A}}$.

Figure 1: Baseline Bayesian models of commitments involving B from B's perspective. Model M_1 represents B's expectations about A satisfying its commitments toward B. Model M_2 represents B's decision to satisfy its commitments toward A.

Trust $\mathsf{T}_{\mathsf{A,B}}$. A's trust in B refers to A's expectation of B to perform a specific task [Castelfranchi and Falcone, 1998; Singh, 2011]. The variable $\mathsf{T}_{\mathsf{A,B}}$ has three values: *low*, *medium*, or *high*.

Emotion E_{A}. We model an agent's emotion as its response to an external or internal event [Friedenberg and Silverman, 2012; Smith and Ellsworth, 1985]. An emotion as E_{A} has three values: *negative* (neg), *neutral* (neu), or *positive* (pos).

4 Baseline Models

We define baseline models and relationships within these models based on existing contributions. Suppose agents Alice (A) and Bob (B) interact with each other. Nodes in the following models are defined for B (A is analogous).

Figure 1 shows models for commitment outcomes. M_1 represents B's current expectations about A satisfying its commitments toward B ($\mathsf{C}^{\mathsf{t}}_{\mathsf{A,B}}$) and M_2 represents B's current decisions to satisfy its commitments toward A ($\mathsf{C}^{\mathsf{t}}_{\mathsf{B,A}}$).

We describe the relationships expressed in these baseline models and how they are justified in the literature. We identify the antecedents and consequents (of commitments and trust) and the conditions (of goals and emotions) as propositional variable, p and q, for relationships where the logical form is directly relevant. We leave ontologies as a basis for abstracting concepts outside our scope.

- $\mathsf{T}^{\mathsf{t-1}}_{\mathsf{B,A}}(p,q) \rightarrow \mathsf{C}^{\mathsf{t}}_{\mathsf{A,B}}(p,q)$ represents that B's trust in A in the previous instant influences B's current expectation of A—i.e., the logical form matters. Indeed, paraphrasing Mayer et al. [1995]: this relationship indicates B's willingness to be vulnerable to the actions of A based on the expectation that A will perform the consequent irrespective of B's ability to monitor or control A.
- $\mathsf{C}^{\mathsf{t-1}}_{\mathsf{B,A}} \rightarrow \mathsf{C}^{\mathsf{t}}_{\mathsf{A,B}}$ represents that B's past decision on the outcomes of commitments toward A influences B's current expectation on the outcomes of commitments from A. That is, B expects something in return; the two commitments

are logically unrelated. This relationship illustrates reciprocity [Hazard and Singh, 2013]. Hence, $C^{t-1}_{A,B} \rightarrow C^{t}_{B,A}$, which represents B's past expectations of the outcomes of commitments from A influences B's current decision on the outcomes of its commitments toward A.

- $T^{t-1}_{B,A} \rightarrow C^{t}_{B,A}$ represents that B's past trust in A influences B's current decision on the outcomes of its commitments toward A. This relationship does not rely upon the logical form. It follows from two relationships: (1) the past trust of B toward A influences B's current expectation on the outcomes of A's commitments and (2) B's past expectations on the outcomes of A's commitments influence B's current decision on the outcomes of B's commitments toward A.
- $E^{t-1}_{B} \rightarrow T^{t}_{B,A}$ represents that B's past emotions influence B's current trust in A [Dunn and Schweitzer, 2005; Antos *et al.*, 2011; Paradeda *et al.*, 2017]. This relationship does not rely upon the logical form of the variables.
- $E^{t-1}_{B} \rightarrow C^{t}_{B,A}$ represents that B's past emotions influence outcomes of B's current commitments toward A. We include this relationship in the baseline since B's emotions influence its trust in A [Dunn and Schweitzer, 2005; Antos *et al.*, 2011; Paradeda *et al.*, 2017] and B's trust in A influences B's decisions to satisfy its commitments (positive outcomes) toward A. This relationship does not rely upon the logical form of the variables.
- $C^{t-1}_{A,B}(p,q) \rightarrow T^{t}_{B,A}(p,q)$ represents that outcomes of A's past commitments toward B influence B's current trust in A. We include this relationship in the baseline based on prior approaches [Kalia *et al.*, 2014; Singh, 2011].
- $G^{t-1}_{B}(q) \rightarrow E^{t}_{B}(q)$ represents that past outcomes of B's goals influence B's current emotions. We include this relationship since Guiraud et al. [2011] and Dastani and Lorini [2012] propose that when B achieves its goals, B's emotions become positive (e.g., joy) and vice versa.
- $G^{t-1}_{B}(q) \rightarrow T^{t}_{B,A}(p,q)$ represents that past outcomes of B's goals influence B's current trust in A. We include this relationship in the baseline assuming that the past outcomes of B's goals influence B's current emotions and B's past emotions influence B's current trust in A.
- $C^{t-1}_{A,B}(p,q) \rightarrow E^{t}_{B}(q)$ represents that the outcomes of A's past commitments toward B influence B's current emotions. We include this relationship since emotions can be responses to the change of state of commitments [Friedenberg and Silverman, 2012; Smith and Ellsworth, 1985].
- $C^{t-1}_{A,B}(p,q) \rightarrow G^{t}_{B}(q)$ represents that the outcomes of A's commitments toward B influence outcomes of B's current goals. We include this relationship in the baseline based on scenarios where when A satisfies its commitments toward B, B achieves its goals [Telang *et al.*, 2019].

We compute the joint distribution for model M_1 as $P(C_{A,B}, C_{B,A}, T_{B,A}, E_B, G_B) = P(C_{A,B} \mid T_{B,A}, C_{B,A})\ P(T_{B,A} \mid E_B, C_{A,B}, G_B)\ P(G_B \mid C_{A,B})\ P(E_B \mid G_B, C_{A,B})$ and for model M_2 as $P(C_{B,A}, C_{A,B}, T_{B,A}, E_B, G_B) = P(C_{B,A} \mid T_{B,A}, C_{A,B}, E_B)\ P(T_{B,A} \mid E_B, C_{A,B}, G_B)\ P(G_B \mid C_{A,B})\ P(E_B \mid G_B, C_{A,B})$.

From the joint distributions for M_1 and M_2, we infer the following conditional dependencies indicating above relationships for the evaluation: (1) $P(C^{t}_{A,B} \mid T^{t-1}_{B,A}, C^{t-1}_{B,A})$, (2) $P(C^{t}_{B,A} \mid T^{t-1}_{B,A}, C^{t-1}_{A,B}, E^{t-1}_{B})$, (3) $P(G^{t}_{B} \mid C^{t-1}_{A,B})$, (4) $P(E^{t}_{B} \mid G^{t-1}_{B}, C^{t-1}_{A,B})$, and (5) $P(T^{t}_{B,A} \mid E^{t-1}_{B}, C^{t-1}_{A,B}, G^{t-1}_{B})$.

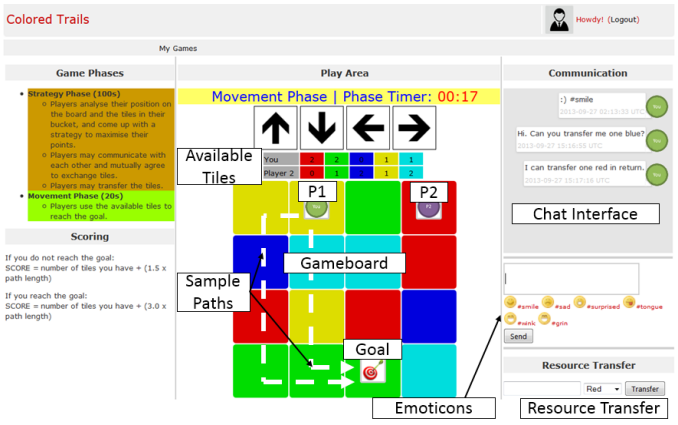


Figure 2: A screenshot of our variant of the Colored Trails game. Left: game phases and scoring instructions; middle: play area; right: communication interface and resource (tiles) transfer panel.

Evaluation criteria. We evaluate the relationships based on three criteria (1) log-likelihood (LL), (2) Akaike Information Criterion (AIC), and (3) Bayesian Information Criterion (BIC) scores, which capture the *goodness of fit* of models to the data. Once we obtain a model, we evaluate its accuracy by computing the area under receiver operating characteristic (ROC) curve (AUC). The curve plots sensitivity versus specificity for a classifier. We obtain an AUC score for each model by averaging three AUC scores obtained by performing three-fold cross-validation respectively on the data for each model. When the data is inadequate for three folds, we consider the AUC score obtained from fewer folds. When we cannot calculate the AUC score at all, we select models based on LL, AIC, and BIC scores. We perform the one-tailed *t-test* on the scores obtained at the 5% significance level.

5 Empirical Study Design

To empirically ground our work, we develop a variant (Figure 2) of [Gal *et al.*, 2010]'s Colored Trails game. Our variant provides a chat interface through which subjects negotiate and exchange tiles and express emotions toward opponents. We associate our variables to the data collected from game play, chats, and surveys provided by subjects. The description below assumes two players, Alice (A) and Bob (B).

Game rules. (1) A subject plays three games with different opponents. (2) Each game consists of five rounds. (3) Each round has a common goal position, and different starting positions for each subject. (4) In each round, subjects are allocated the same number but a different set (randomly selected) of colored tiles. (5) Subjects can communicate with their opponents via a chat interface, in which they can negotiate to transfer tiles to each other. (6) At the beginning of a game and at the end of each round, each subject fills a survey. (7) Let n be the number of tiles left unused and u be the number of tiles left used. The score for a subject who (a) does not

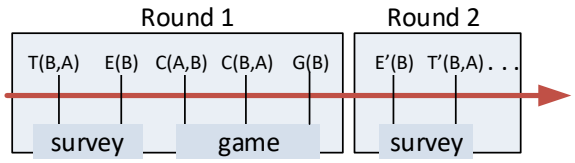


Figure 3: Timeline of our data acquisition for B. (Similarly for A.)

reach his or her goal is $n + 1.5 \times u$ and (b) reaches his or her goal is $n + 3.0 \times u$.

Mapping concepts with model. We describe the mapping between the game and concepts for B. The mapping for A is the same. (1) *Goal*. B's goal is to reach the *goal* position. (2) *Commitment*. During the game, A can create a commitment by agreeing (through chat) to transfer the specified tiles to B, usually in exchange for some tiles. If A provides the tiles, A satisfies his or her commitment toward B, otherwise A violates it. (4) *Emotion*. B's emotion changes as the games progress and may spill over from one game to the next. We determine B's emotion from B's survey response. (5) *Trust*. We determine B's trust in A from B's survey response.

Subjects. We recruited 30 (25 male; 5 female) subjects, all students in Computer Science. We offered a payment of 10–20 USD each, depending on success in the game. We obtained IRB approval, and informed consent from subjects.

Surveys. (1) At the start of each game and at the end of each round, we asked a subject to record his or her trust for an opponent and an emotion on a five-point scale (very negative, negative, neutral, positive, very positive). We mapped the responses to our trust ($\mathsf{T}_{\mathsf{A,B}}$) and emotion ($\mathsf{E}_{\mathsf{A}}$) variables, respectively. For analysis, we converted the above five-point scale into a three-point scale by merging very negative and negative responses and very positive and positive responses.

Threats to validity and mitigation. We discuss some important threats to validity to our empirical study. First, the subjects could know each other from before. Thus, we anonymize and separate them to prevent them from knowing who their opponents are or sending any visual or auditory signals. Second, a subject could be an expert in the domain. We select subjects who do not have prior experience in our topic. Third, a subject could produce a strong negative emotional response from the outcomes of a few rounds and might decide to leave the game. We limit their emotional response by setting the game with lower stakes for success or failure.

6 Dataset and Processing

We collected 450 rows of data (30 subjects $\times$ 30 games $\times$ 5 rounds per game), including their survey forms, and whatever chat messages they exchanged. From subjects' interactions, we manually analyzed messages exchanged by the subjects to identify commitments ($\mathsf{C}_{\mathsf{B,A}}$) and ($\mathsf{C}_{\mathsf{A,B}}$) and their outcomes. For each commitment, we add preceding emotions (E_{A}), trust ($\mathsf{T}_{\mathsf{A,B}}$), and goals ($\mathsf{G}_{\mathsf{A}}$). Table 1 shows the data distribution for each variable. Note that the count of commitments satisfied or violated is much smaller than the number of rounds played, indicating that in several rounds, either the players could achieve their goals without cooperation or that their negotiations failed.

Variable	**Value**$_1$	**Value**$_2$	**Value**$_3$
Commitments	sat (119)	vio (16)	
Goals	ach (231)	fai (219)	
Trust	low (95)	med (224)	high (131)
Emotions	neg (85)	neu (183)	pos (182)

Table 1: Distributions of values of variables in the data.

	$\mathsf{T}^1_{\mathsf{B,A}}$	$\mathsf{E}^1_{\mathsf{B}}$	$\mathsf{C}^1_{\mathsf{A,B}}$	$\mathsf{G}^1_{\mathsf{B}}$	$\mathsf{C}^1_{\mathsf{B,A}}$
$\mathsf{T}^1_{\mathsf{B,A}}$	1.00	0.40	0.02	0.10	0.20
$\mathsf{E}^1_{\mathsf{B}}$	0.40	1.00	0.05	−0.01	−0.20
$\mathsf{C}^1_{\mathsf{A,B}}$	0.02	0.10	1.00	0.20	−0.20
$\mathsf{G}^1_{\mathsf{B}}$	0.10	−0.01	0.20	1.00	0.10
$\mathsf{C}^1_{\mathsf{B,A}}$	0.20	−0.10	−0.20	0.10	1.00

Table 2: Correlations of variables in one slice. (Superscript = round.)

Random variables. We represent the models in terms of a Dynamic Bayesian Network (DBN) with two time slices, $t-1$ and t. To do so, we mapped each round in a game to a time slice. For example, we mapped time slice $t-1$ to Round 1 and the next time slice t to Round 2. In each round, for each player, we observed five discrete random variables. As shown in Figure 3, first, B records its trust in A, $\mathsf{T}_{\mathsf{B,A}}$ and emotion, E_{B}. Then, in the game, in each round, the following variables were observed for B: outcomes of A's commitments toward B ($\mathsf{C}_{\mathsf{A,B}}$), outcomes of B's commitments toward A ($\mathsf{C}_{\mathsf{B,A}}$), and outcomes of B's goals (G_{B}).

Correlations. Before we created a DBN, we determined Pearson's correlation coefficients R between the random variables observed during the game. Table 2 shows the correlations between variables observed in time slice t whereas Table 3 shows the correlations between variables observed in time slice $t-1$ and the variables observed in time slice t. We create these tables to observe if random variables have static or dynamic (causal) relationships.

Observations. From the correlation coefficients, we obtain the overall relationships between random variables. Based on the prior work [Dunn and Schweitzer, 2005; Steunebrink *et al.*, 2007; Guiraud *et al.*, 2011], we only consider positive correlations between the variables. For example, Dunn and Schweitzer [2005] observe that negative emotions decrease trust and vice versa. Steunebrink et al. [2007] suggest

	$\mathsf{T}^2_{\mathsf{B,A}}$	$\mathsf{E}^2_{\mathsf{B}}$	$\mathsf{C}^2_{\mathsf{A,B}}$	$\mathsf{G}^2_{\mathsf{B}}$	$\mathsf{C}^2_{\mathsf{B,A}}$
$\mathsf{T}^1_{\mathsf{B,A}}$	0.60	0.30	0.30	0.04	0.20
$\mathsf{E}^1_{\mathsf{B}}$	0.30	0.20	0.10	0.01	0.10
$\mathsf{C}^1_{\mathsf{A,B}}$	0.20	0.10	0.40	−0.02	0.30
$\mathsf{G}^1_{\mathsf{B}}$	0.40	0.50	−0.03	−0.10	NaN
$\mathsf{C}^1_{\mathsf{B,A}}$	0.20	0.10	0.30	0.10	0.20

Table 3: Correlating variables across slices. (Superscript = round.)

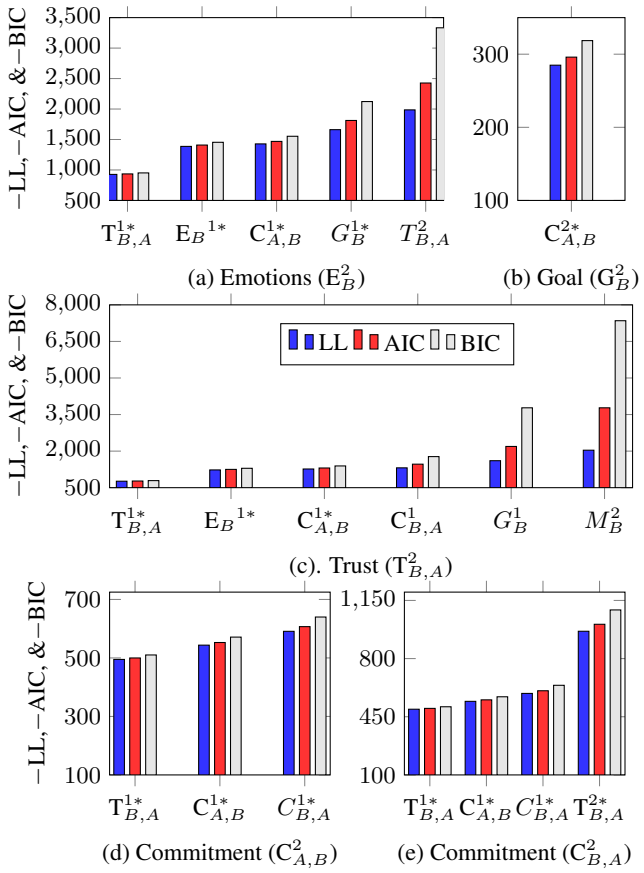


Figure 4: Negated LL, AIC, and BIC scores for all possible Bayesian models obtained from correlations, (*p<0.05).

goal achievement brings positive emotions such as joy. In Tables 2 and 3, positive correlations vary from 0.01 to 0.6. We consider correlations where R is greater than or equal to 0.1. An R of 0.1 may indicate a weak correlation. However, such correlations are common in psychology [Chiaburu *et al.*, 2011]. Based on the observed correlations, we incrementally created Bayesian models for emotions, goals, trust, and commitments by adding one variable at a time as input to a target variable ($T^2_{B,A}$, E^2_B, $C^2_{A,B}$, $C^2_{B,A}$, G^2_B). For each such incremental model, we calculated the LL, AIC, and BIC scores—and show them in Figure 4.

7 Producing Refined Models

We evaluate the existing relationships in the baseline models (M_1 and M_2 shown in Figure 1) using the log-likelihood (LL, AIC, and BIC) scores. In this process, we discover new relationships that were not present in the baseline models. We include only the relationships that produce significant changes and derive the new models M'_1 and M'_2 as shown in Figure 5.

Rel$_1$. P ($C^t_{A,B} \mid T^{t-1}_{B,A}$, $C^{t-1}_{B,A}$) represents a player's expectations about its opponent satisfying a commitment toward the player. From the LL, AIC, and BIC scores, we found that when $T^{t-1}_{B,A}$, $C^{t-1}_{A,B}$, and $C^{t-1}_{B,A}$ were incorporated in the model, the changes in scores were significant with p-values below 0.01 in each case. This suggests that the existing relationship hold. In addition, we learned a new relationship $C^{t-1}_{A,B} \rightarrow C^t_{A,B}$ from the data. We also computed the AUC scores for $C^t_{A,B} \mid T^{t-1}_{B,A}$ (0.69), $C^t_{A,B} \mid C^{t-1}_{B,A}$ (0.68), and $C^t_{A,B} \mid C^{t-1}_{A,B}$ (0.71) suggesting the past satisfaction of expectations has the highest influence on $C^t_{A,B}$.

Rel$_2$. P ($C^t_{B,A} \mid T^{t-1}_{B,A}$, $C^{t-1}_{A,B}$, E^{t-1}_B) represents a player's decision to satisfy its commitments. From the LL, AIC, and BIC scores, we found that when $T^{t-1}_{B,A}$, $C^{t-1}_{A,B}$, $C^{t-1}_{B,A}$, and $T^t_{B,A}$ were incorporated into the model, the changes in scores, shown in Figure 4(d), were significant (with p-values of 0.00, 0.00, 0.01, and 0.00, respectively). This suggests the relationships $T^{t-1}_{B,A} \rightarrow C^t_{B,A}$ and $C^{t-1}_{A,B} \rightarrow C^t_{B,A}$ hold whereas $E^{t-1}_B \rightarrow C^t_{B,A}$ does not hold. In addition, we learned two new relationships: $C^{t-1}_{B,A} \rightarrow C^t_{B,A}$ and $T^t_{B,A} \rightarrow C^t_{B,A}$. We compare the AUC scores for $C^t_{B,A} \mid T^{t-1}_{B,A}$ (0.56), $C^t_{B,A} \mid C^{t-1}_{A,B}$ (0.71), $C^t_{B,A} \mid C^{t-1}_{B,A}$ (0.68), and $C^t_{B,A} \mid T^t_{B,A}$ (0.71). This comparison suggests $C^{t-1}_{A,B}$ and $T^t_{B,A}$ have the highest influence on $C^t_{B,A}$.

Rel$_3$. P($G^t_B \mid C^{t-1}_{A,B}$) represents the goals of a player. From the LL, AIC, BIC scores for each such model, we found that when $C^t_{A,B}$ was added to the model, the change in the scores for $C^t_{A,B}$ was significant (with a p-value of 0.01). This suggests that the existing relationship does not hold indicating the outcomes of past commitments do no influence the current outcome of a goal. The new relationship we obtain is $C^t_{A,B} \rightarrow G^t_B$ and its AUC score for $G^t_B \mid G^t_B$ is 0.56.

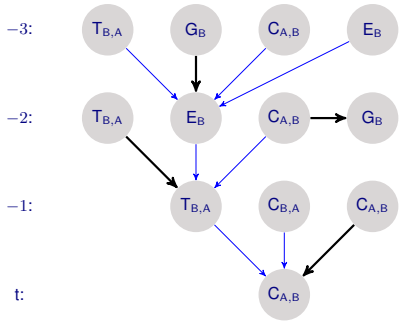
Rel$_4$. P($E^t_B \mid G^{t-1}_B$, $C^{t-1}_{A,B}$) represents the emotions of a player. From the LL, AIC, and BIC scores, we found that when trust $T^{t-1}_{B,A}$, emotions E^{t-1}_B, commitments $C^{t-1}_{A,B}$, and goals G^{t-1}_B were incorporated in the model, the changes in the scores were significant with p-values of 0.00, 0.00, 0.03, and 0.02, respectively). This means that the baseline relationships $G^{t-1}_B \rightarrow E^t_B$ and $C^{t-1}_{A,B} \rightarrow E^t_B$ hold. In addition, we learned two new relationships $E^{t-1}_B \rightarrow E^t_B$ and $T^{t-1}_{B,A} \rightarrow E^t_B$. We obtain the AUC scores for $E^t_B \mid G^{t-1}_B$ (0.69), $E^t_B \mid C^{t-1}_{A,B}$ (0.5), $E^t_B \mid E^{t-1}_B$ (0.56), and $E^t_B \mid T^1_{B,A}$ (0.59) to find that G^{t-1}_B has the highest influence on E^t_B followed by E^{t-1}_B and $T^{t-1}_{B,A}$.

Rel$_5$. P ($T^t_{B,A} \mid E^{t-1}_B$, $C^{t-1}_{A,B}$, G^{t-1}_B) represents trust of a player for another player. From the LL, AIC, and BIC scores, we found that when $T^t_{B,A}$, E^{t-1}_B, and $C^{t-1}_{A,B}$ were incorporated in the model, the changes in the scores were significant with p-values of 0.00, 0.00, and 0.03, respectively. This suggests that the relationships $C^{t-1}_{A,B} \rightarrow T^t_{B,A}$ and $E^{t-1}_B \rightarrow T^t_{B,A}$ hold and $G^{t-1}_B \rightarrow T^t_{B,A}$ does not hold. In addition, we learned a new relationship $T^{t-1}_{B,A} \rightarrow T^t_{B,A}$. We obtain the AUC scores for $T^t_{B,A} \mid T^{t-1}_{B,A}$ (0.74), $T^t_{B,A} \mid E^{t-1}_B$ (0.63), and $T^t_{B,A} \mid C^{t-1}_{A,B}$ (0.47) to find that $T^{t-1}_{B,A}$ has the highest influence on $T^t_{B,A}$.

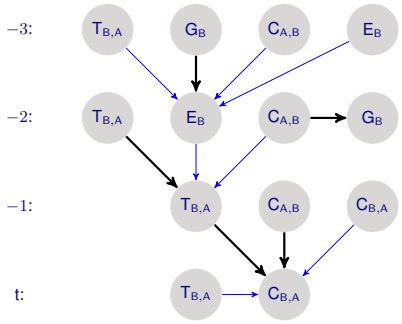
Table 4 compares the relationships in the baseline models and the new models. Figure 5 shows the new models, M'_1 and M'_2, produced by combining the new relationships

Baseline Relationships	Changes	New Relationships
Rel$_1$: P ($C^t_{A,B}$ | $T^{t-1}_{B,A}$, $C^{t-1}_{B,A}$)	added: $C^{t-1}_{A,B}$	P ($C^t_{A,B}$ | $T^{t-1}_{B,A}$, $C^{t-1}_{B,A}$,$C^{t-1}_{A,B}$)
Rel$_2$: P ($C^t_{B,A}$ | $T^{t-1}_{B,A}$, $C^{t-1}_{A,B}$, E^{t-1}_B)	removed: E^{t-1}_B; $C^{t-1}_{B,A}$, $T^t_{B,A}$	P ($C^t_{B,A}$ | $T^{t-1}_{B,A}$, $C^{t-1}_{A,B}$, $C^{t-1}_{B,A}$, $T^t_{B,A}$)
Rel$_3$: P (G^t_B |$C^{t-1}_{A,B}$)	removed: $C^{t-1}_{A,B}$; added: $C^t_{A,B}$	P (G^t_B | $C^t_{A,B}$)
Rel$_4$: P (E^t_B | G^{t-1}_B, $C^{t-1}_{A,B}$)	added: E^{t-1}_B, $T^{t-1}_{B,A}$	P (E^t_B | G^{t-1}_B, $C^{t-1}_{A,B}$, E^{t-1}_B, $T^{t-1}_{B,A}$)
Rel$_5$: P ($T^t_{B,A}$ | E^{t-1}_B, $C^{t-1}_{A,B}$, G^{t-1}_B)	removed: G^1_B; added: $T^{t-1}_{B,A}$	P ($T^t_{B,A}$ | E^{t-1}_B, $C^{t-1}_{A,B}$, $T^{t-1}_{B,A}$)

Table 4: Summary of comparisons between the baseline and our proposed relationships based on LL, AIC, and BIC scores.



(a) Model M'_1 for $C_{A,B}$.



(b) Model M'_2 for $C_{B,A}$.

Figure 5: Resulting Bayesian models of commitments involving B from B's perspective. Model M'_1 represents B's expectations about A satisfying its commitments toward B. Model M'_2 represents B's decision to satisfy its commitments toward A. Bold edges indicate the strongest relationships—in one case two edges have same scores. Note that there are five unique strongest relationships.

for the respective target variables, and highlights the relationships that yield the highest AUC scores of any relationship with the same conclusion.

8 Discussion and Conclusion

We initially construct two baseline models, M_1 and M_2, based on prior work that capture the outcome of commitments based on emotions along with trust and goals. We evaluate these models using the LL, AIC, BIC, and AUC scores based on data from human subjects. Accordingly, we construct two new models, M'_1 and M'_2, by adding and removing relationships based on their empirical backing. We identify the following relationships as among those most strongly supported.

(1) *The outcomes of an agent's past commitment to a second agent influence the second agent's current expectations of the outcomes of a commitment from the first.* **(2)** *The outcomes of an agent's past commitment to another agent influence the second agent's current decision on the outcome of its commitments to the first.* **(3)** *An agent's past trust in a second agent influences the agent's current decision on the outcome of its commitments to the second.* **(4)** *The outcomes of an agent's past goals strongly influence its current emotions.* **(5)** *An agent's past trust in another agent strongly influences its current trust in the second agent.*

Real-world applications. Empirically grounded Bayesian models promise to support real-world applications, such as the following. (1) We can compute trust between team members based on their commitments, trust, and emotions, to recommend optimal team configurations [Kalia *et al.*, 2017; Liemhetcharat and Veloso, 2012]. (2) We can enhance Mayer et al.'s [1995] Trust Antecedent Framework to incorporate emotions to recommend the amount of risk team leaders can take while assigning important tasks to team members. (3) We can use the dependency of emotions on trust to filter out dishonest advisors more accurately than by considering trust alone [Irissappane and Zhang, 2017]. We can leverage this dependency to build an agent that discovers valuable messages via trustworthy peers in a social network [Sardana *et al.*, 2017]. (4) We can build help-desk cognitive assistants [Telang *et al.*, 2018] who seek to solve customers' problems. An assistant would benefit from interpreting customers' emotions and trust to engage effectively with customers, including by prioritizing between problems that a customer brings up.

Future work. We plan to evaluate our models on large-scale communication data to gauge their effectiveness and generality. Doing so would involve sophisticated text analysis to extract norms, trust, goals, and emotions. We can enhance the models proposed in this paper to incorporate intensity of emotions and trust. We can include additional emotions such as hope, joy, and sorrow to bring forth more interesting insights into the existing models for predicting norm outcomes.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This research was partially supported by the US DoD through the Science of Security Lablet (SoSL) at NCSU, a US ARL ORISE Fellowship, and by IBM Research.

References

[Ajmeri *et al.*, 2018] N. Ajmeri, H. Guo, P.K. Murukannaiah, M.P. Singh. Robust norm emergence by revealing and reasoning about context: Socially intelligent agents for enhancing privacy. *IJCAI*, pp. 28–34, 2018.

[Antos *et al.*, 2011] D. Antos, C. de Melo, J. Gratch, B. Grosz. The influence of emotion expression on perceptions of trustworthiness in negotiation. *AAAI*, 2011.

[Arnold, 1960] M. B. Arnold. *Emotion and Personality*. Columbia University, New York, 1960.

[Brooks *et al.*, 2011] L. Brooks, W. Iba, S. Sen. Modeling the emergence and convergence of norms. *IJCAI*, 2011.

[Castelfranchi and Falcone, 1998] C. Castelfranchi, R. Falcone. Principles of trust for MAS. *ICMAS*, 72–79. 1998.

[Chiaburu *et al.*, 2011] D.S. Chiaburu, I.-S. Oh, C.M. Berry, N. Li, R.G. Gardner. The five-factor model of personality traits and organizational citizenship behaviors. *J. Applied Psych.*, 96(6):1140–1166, 2011.

[Dastani and Lorini, 2012] M. Dastani, E. Lorini. A logic of emotions. *AAMAS*, pp. 1133–1140, 2012.

[de Melo *et al.*, 2012] C.M. de Melo, P. Carnevale, S. Read, D. Antos, J. Gratch. Bayesian model of the social effects of emotion in decision-making in multiagent systems. *AAMAS*, pp. 55–62, 2012.

[Dunn and Schweitzer, 2005] J.R. Dunn, M.E. Schweitzer. Feeling and believing: The influence of emotion on trust. *J. Personality & Social Psych.*, 88(5):736–748, 2005.

[Forgas, 1995] J.P. Forgas. Mood and judgment: The Affect Infusion Model (AIM). *Psych. Bull.*, 117(1):39–66, 1995.

[Friedenberg and Silverman, 2012] J. Friedenberg, G. Silverman. *Cognitive Science*. Sage, 2nd edition, 2012.

[Gal *et al.*, 2010] Y. Gal, B. Grosz, S. Kraus, A. Pfeffer, S. Shieber. Agent decision-making in open-mixed networks. *Artificial Intelligence*, 174(18):1460–1480, 2010.

[Guiraud *et al.*, 2011] N. Guiraud, D. Longin, E. Lorini, S. Pesty, J. Rivière. The face of emotions *AAMAS*, 2011.

[Hazard and Singh, 2013] C.J. Hazard, M.P. Singh. Macau: A basis for evaluating reputation systems. *IJCAI*, 2013.

[Hoegen *et al.*, 2017] R. Hoegen, G. Stratou, J. Gratch. Incorporating emotion perception into opponent modeling for social dilemmas. *AAMAS*, pp. 801–809, 2017.

[Irissappane and Zhang, 2017] A.A. Irissappane, J. Zhang. Filtering unfair ratings from dishonest advisors in multi-criteria e-markets *JAAMAS*, 31(1):36–65, 2017.

[Kalia *et al.*, 2014] A.K. Kalia, Z. Zhang, M.P. Singh. Estimating trust from agents' interactions via commitments. *ECAI*, pp. 1043–1044, 2014.

[Kalia *et al.*, 2017] A. Kalia, N. Buchler, A. DeCostanza, M. Singh. Computing team performance measures from the structure and content of broadcast collaborative communications. *IEEE Trans. Comp. Soc. Syst.*, 4(2):26–39, 2017.

[Lallé *et al.*, 2018] S. Lallé, C. Conati, R. Azevedo. Prediction of student achievement goals and emotion valence during interaction with pedagogical agents. *AAMAS*, pp. 1222–1231, 2018.

[Lazarus, 1966] R.S. Lazarus. *Psychological Stress and the Coping Process*. McGraw-Hill, New York, 1966.

[Liemhetcharat and Veloso, 2012] S. Liemhetcharat, M. Veloso. Modeling and learning synergy for team formation with heterogeneous agents. *AAMAS*, 2012.

[Lorini and Schwarzentruber, 2009] E. Lorini, F. Schwarzentruber. A logic for reasoning about counterfactual emotions. *IJCAI*, pp. 867–872, 2009.

[Mahmoud *et al.*, 2016] S. Mahmoud, S. Miles, M. Luck. Cooperation emergence under resource-constrained peer punishment. *AAMAS*, pp. 900–908, 2016.

[Mayer *et al.*, 1995] R.C. Mayer, J.H. Davis, F.D. Schoorman. An integrative model of organizational trust. *The Academy of Management Review*, 20(3):709–734, 1995.

[Nardin *et al.*, 2016] L. Nardin, T. Balke-Visser, N. Ajmeri, A. Kalia, J. Sichman, M. Singh. Classifying sanctions and designing a conceptual sanctioning process model for socio-technical systems. *Know Eng Rev*, 31:142–166, 2016.

[Ortony *et al.*, 1988] A. Ortony, G.L. Clore, A. Collins. *The Cognitive Structure of Emotions*. Camb. U Press, 1988.

[Paradeda *et al.*, 2017] R. Paradeda, M. Hashemian, C. Guerra, R. Prada, J. Dias, A. Paiva. Fides: How emotions and small talks may influence trust in an embodied vs. non-embodied robot. *AAMAS*, pp. 1673–1675, 2017.

[Russell, 1980] J.A. Russell. A circumplex model of affect. *J. Personality & Social Psych.*, 39(6):1161–1178, 1980.

[Sardana *et al.*, 2017] N. Sardana, R. Cohen, J. Zhang, S. Chen. A Bayesian multiagent trust model for social networks. *IEEE Trans Comp Soc Syst*, 5(4):995–1008, 2018.

[Singh, 1999] M.P. Singh. An ontology for commitments in multiagent systems. *AI & Law*, 7(1):97–113, 1999.

[Singh, 2011] M.P. Singh. Trust as Dependence: A Logical Approach. *AAMAS*, pp. 863–870, 2011.

[Smith and Ellsworth, 1985] C.A. Smith, P.C. Ellsworth. Patterns of cognitive appraisal in emotion. *J. Personality and Social Psych.*, 48(4):813–838, 1985.

[Steunebrink *et al.*, 2007] B.R. Steunebrink, M. Dastani, J.-J. Ch. Meyer. A logic of emotions for intelligent agents. *AAAI*, pp. 142–147, 2007.

[Telang *et al.*, 2019] P.R. Telang, M.P. Singh, N. Yorke-Smith. A coupled operational semantics for goals and commitments. *JAIR*, 65(2):31–85, 2019.

[Telang *et al.*, 2018] P.R. Telang, A.K. Kalia, M. Vukovic, R. Pandita, M.P Singh A conceptual framework for engineering chatbots. *IEEE Internet Computing.*, 22(6):54–59, 2018.