

# The Interplay of Emotions and Norms in Multiagent Systems

Anup K. Kalia<sup>1</sup>, Nirav Ajmeri<sup>2</sup>, Kevin Chan<sup>3</sup>, Jin-Hee Cho<sup>4</sup>, Sibel Adali<sup>5</sup>, Munindar P. Singh<sup>2</sup>

IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA<sup>1</sup>

US Army Research Lab, Aberdeen Proving Ground, MD 21005, USA<sup>3</sup>

Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA<sup>4</sup>

Rensselaer Polytechnic Institute, Troy, NY 12180, USA<sup>5</sup>

North Carolina State University, Raleigh, NC 27695, USA<sup>2</sup>

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 baseline models to discover new insights into the existing 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 affected by whether its goals are achieved based on norms 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. (Alice may additionally create a commitment norm with Bob to pay him for the completed task, though this does not feature in our scenario.) Bob agrees, meaning he commits to Alice to perform the specified 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.*

**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. This might lead Alice to lose trust in Bob. As a result, both Alice and Bob might feel discouraged from interacting further.*

Examples 1 and 2 suggest that Alice's decision to commit to Bob depends on the outcomes of her prior interactions with him. 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. For example, Nardin *et al.* [2016] suggest self-directed sanctions (e.g., guilt and trust) and other-directed sanctions (e.g., gossip and praise). Thus, it is important to consider norm outcomes with respect to emotions, trust, and goals. Existing works capture specific relationships among these variables: emotions and trust [Dunn and Schweitzer, 2005; Antos *et al.*, 2011; Paradedda *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.

To develop 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 and effectively study the relationships among the above variables. 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 describe the 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 Paradedda 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 based on decision-making 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 Schwarzen-truber [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 and most importantly neither provide nor evaluate computational models for predicting norm outcomes. Kalia et al. [2014] provide some experimental evaluations, however, the model they consider 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. 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  $C_{A,B}$ .** A commitment  $C_{A,B}$  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  $G_A$  can have a binary value, *achieved* (ach) or *failed* (fai).

**Trust  $T_{A,B}$ .** 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  $T_{A,B}$  can have one of the three possible 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  $E_A$  can have one of the three possible 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 ( $C_{A,B}^t$ ) and  $M_2$  represents B’s current decisions to satisfy its commitments toward A ( $C_{B,A}^t$ ).

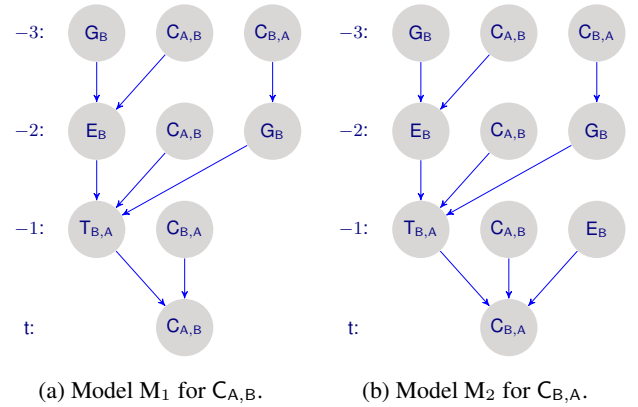


Figure 1: Baseline Bayesian models of commitments involving B from B’s perspective. Model  $M_1$  represents B’s expectations about A satisfying its commitments toward B. Model  $M_2$  represents B’s decision to satisfy its commitments toward A.

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.

- $T_{B,A}^{t-1}(p, q) \rightarrow C_{A,B}^t(p, q)$  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.
- $C_{B,A}^{t-1} \rightarrow C_{A,B}^t$  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—i.e., an illustration of reciprocity [Hazard and Singh, 2013]. Hence,  $C_{A,B}^{t-1} \rightarrow C_{B,A}^t$ , 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.

- $T_{B,A}^{t-1} \rightarrow C_{B,A}^t$  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.
- $E_B^{t-1} \rightarrow T_{B,A}^t$  represents that B's past emotions influence B's current trust in A [Dunn and Schweitzer, 2005; Antos *et al.*, 2011; Paradedá *et al.*, 2017]. This relationship does not rely upon the logical form of the variables.
- $E_B^{t-1} \rightarrow C_{B,A}^t$  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; Paradedá *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.
- $C_{A,B}^{t-1}(p, q) \rightarrow T_{B,A}^t(p, q)$  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].
- $G_B^{t-1}(q) \rightarrow E_B^t(q)$  represents that past outcomes of B's goals influence B's current emotions. We include the relationship in the baseline 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.
- $G_B^{t-1}(q) \rightarrow T_{B,A}^t(p, q)$  represents that past outcomes of B's goals influence B's current trust in A. We include the 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.
- $C_{A,B}^{t-1}(p, q) \rightarrow E_B^t(q)$  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].
- $C_{A,B}^{t-1}(p, q) \rightarrow G_B^t(q)$  represents that the outcomes of A's commitments toward B influence outcomes of B's current goals. We include the 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  $P(C_{A,B}, C_{B,A}, T_{B,A}, E_B, G_B) = P(C_{A,B} | T_{B,A}, C_{B,A}) P(T_{B,A} | E_B, C_{A,B}, G_B) P(G_B | C_{A,B}) P(E_B | G_B, C_{A,B})$  and for model  $M_2$  as  $P(C_{B,A}, C_{A,B}, T_{B,A}, E_B, G_B) = P(C_{B,A} | T_{B,A}, C_{A,B}, E_B) P(T_{B,A} | E_B, C_{A,B}, G_B) P(G_B | C_{A,B}) P(E_B | G_B, C_{A,B})$ .

From the joint distributions for  $M_1$  and  $M_2$ , we infer the following conditional dependencies indicating above relationships for the evaluation: (1)  $P(C_{A,B}^t | T_{B,A}^{t-1}, C_{B,A}^{t-1})$ , (2)  $P(C_{B,A}^t | T_{B,A}^{t-1}, C_{A,B}^{t-1}, E_B^{t-1})$ , (3)  $P(G_B^t | C_{A,B}^{t-1})$ , (4)  $P(E_B^t | G_B^{t-1}, C_{A,B}^{t-1})$ , and (5)  $P(T_{B,A}^t | E_B^{t-1}, C_{A,B}^{t-1}, G_B^{t-1})$ .

**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).

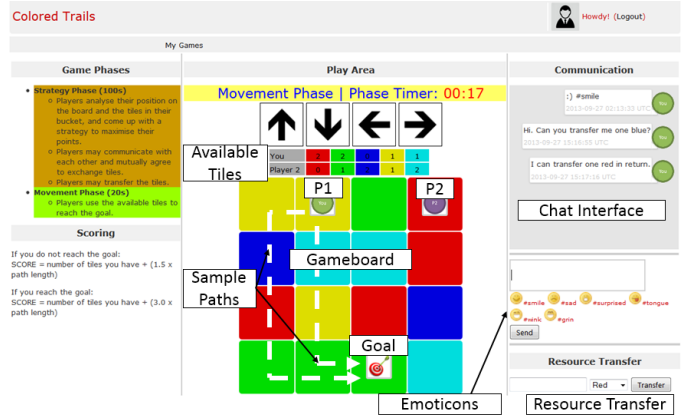


Figure 2: A screenshot of our variant of the Colored Trails game. Left: game phases and scoring instructions; middle: play area; right: communication interface and resource (tiles) transfer panel.

**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 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 commit-

ment 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 ( $T_{A,B}$ ) and emotion ( $E_A$ ) 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.

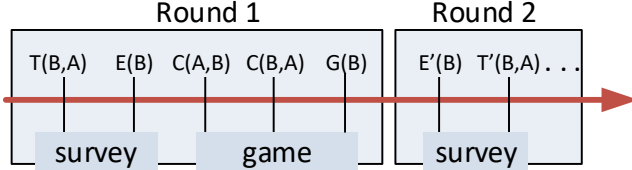


Figure 3: Timeline of our data acquisition for B. (Similarly for A.)

**Threats to Validity and Mitigation** The empirical study has several threats to validity. First, the subjects could know each other from before. We mitigate the threat by anonymizing and separating them to prevent them from knowing about their opponents or sending any visual or auditory signals. Second, a subject could be an expert in the domain. We mitigate this threat by selecting subjects who do not have prior experiences in our topic. Third, a subject could emit strong negative emotional response from the outcomes of every round in the game that might discourage the subject 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 ( $C_{B,A}$ ) and ( $C_{A,B}$ ) and their outcomes. For each commitment, we add preceding emotions ( $E_A$ ), trust ( $T_{A,B}$ ), and goals ( $G_A$ ). 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.

**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

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

Variable	Value <sub>1</sub>	Value <sub>2</sub>	Value <sub>3</sub>
Commitments	sat (119)	vio (16)	
Goals	ach (231)	fai (219)	
Trust	low (95)	med (224)	high (131)
Emotions	neg (85)	neu (183)	pos (182)

player, we observed five discrete random variables. As shown in Figure 3, first, B records its trust in A,  $T_{B,A}$  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 ( $C_{A,B}$ ), outcomes of B’s commitments toward A ( $C_{B,A}$ ), 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.

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

	$T_{B,A}^1$	$E_B^1$	$C_{A,B}^1$	$G_B^1$	$C_{B,A}^1$
$T_{B,A}^1$	1.00	0.40	0.02	0.10	0.20
$E_B^1$	0.40	1.00	0.05	−0.01	−0.20
$C_{A,B}^1$	0.02	0.10	1.00	0.20	−0.20
$G_B^1$	0.10	−0.01	0.20	1.00	0.10
$C_{B,A}^1$	0.20	−0.10	−0.20	0.10	1.00

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

	$T_{B,A}^2$	$E_B^2$	$C_{A,B}^2$	$G_B^2$	$C_{B,A}^2$
$T_{B,A}^1$	0.60	0.30	0.30	0.04	0.20
$E_B^1$	0.30	0.20	0.10	0.01	0.10
$C_{A,B}^1$	0.20	0.10	0.40	−0.02	0.30
$G_B^1$	0.40	0.50	−0.03	−0.10	NaN
$C_{B,A}^1$	0.20	0.10	0.30	0.10	0.20

**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 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 ( $T_{B,A}^2$ ,  $E_B^2$ ,  $C_{A,B}^2$ ,  $C_{B,A}^2$ ,  $G_B^2$ ). For each such incremental model, we calculated the LL, AIC, and BIC scores—and show them in Figure 4.



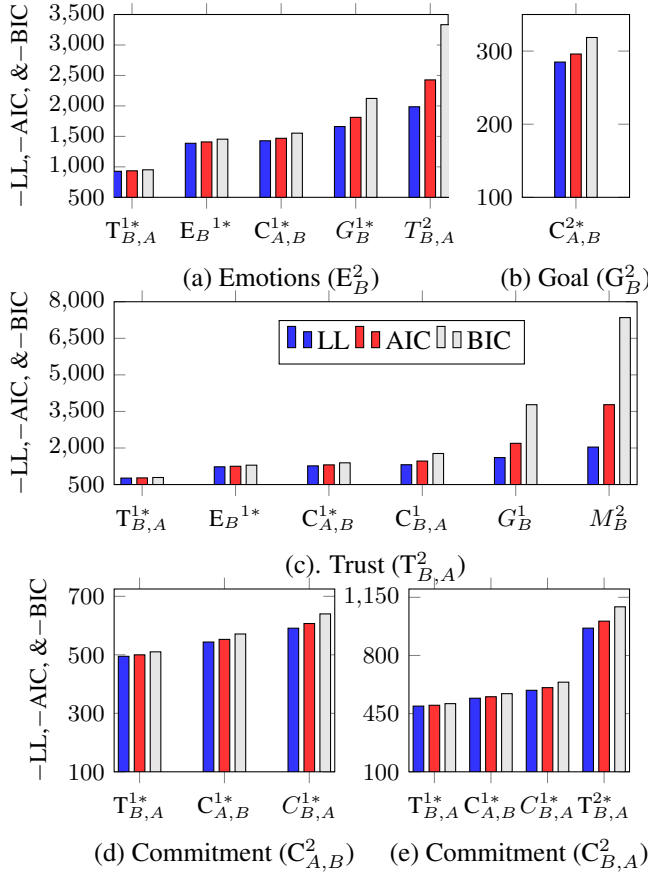


Figure 4: Negated LL, AIC, and BIC scores for all possible Bayesian models obtained from correlations, (\* $p < 0.05$ ).

## 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 Dynamic Bayesian models  $M'_1$  and  $M'_2$  as shown in Figure 5.

**Rel<sub>1</sub>**  $P(C_{A,B}^t | T_{B,A}^{t-1}, C_{B,A}^{t-1})$  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  $T_{B,A}^{t-1}$ ,  $C_{A,B}^{t-1}$ , and  $C_{B,A}^{t-1}$  were incorporated in the model, the changes in scores were significant with p-values of 0.00, 0.00, and 0.00, respectively. This suggests that the existing relationship hold. In addition, we learned a new relationship  $C_{A,B}^{t-1} \rightarrow C_{A,B}^t$  from the data. We also computed the AUC scores for  $C_{A,B}^t | T_{B,A}^{t-1}$  (0.69),  $C_{A,B}^t | C_{B,A}^{t-1}$  (0.68), and  $C_{A,B}^t | C_{A,B}^{t-1}$  (0.71) suggesting the past satisfaction of expectations has the highest influence on  $C_{A,B}^t$ .

**Rel<sub>2</sub>**  $P(C_{B,A}^t | T_{B,A}^{t-1}, C_{A,B}^{t-1}, E_B^{t-1})$  represents a player's decision to satisfy its commitments. From the LL, AIC, and BIC scores, we found that when  $T_{B,A}^{t-1}$ ,  $C_{A,B}^{t-1}$ ,  $C_{B,A}^{t-1}$ , and  $E_B^{t-1}$  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  $T_{B,A}^{t-1} \rightarrow C_{B,A}^t$  and  $C_{A,B}^{t-1} \rightarrow C_{B,A}^t$  hold whereas  $E_B^{t-1} \rightarrow C_{B,A}^t$  does not hold. In addition, we learned two new relationships:  $C_{B,A}^{t-1} \rightarrow C_{B,A}^t$  and  $T_{B,A}^t \rightarrow C_{B,A}^t$ . We compare the AUC scores for  $C_{B,A}^t | T_{B,A}^{t-1}$  (0.56),  $C_{B,A}^t | C_{A,B}^{t-1}$  (0.71),  $C_{B,A}^t | C_{B,A}^{t-1}$  (0.68), and  $C_{B,A}^t | T_{B,A}^t$  (0.71). This comparison suggests  $C_{A,B}^{t-1}$  and  $T_{B,A}^t$  have the highest influence on  $C_{B,A}^t$ .

**Rel<sub>3</sub>**  $P(G_B^t | C_{A,B}^{t-1})$  represents the goals of a player. From the LL, AIC, BIC scores for each such model, we found that when  $C_{A,B}^t$  was added to the model, the change in the scores for  $C_{A,B}^t$  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  $C_{A,B}^t \rightarrow G_B^t$  and its AUC score for  $G_B^t | G_B^t$  is 0.56.

**Rel<sub>4</sub>**  $P(E_B^t | G_B^{t-1}, C_{A,B}^{t-1})$  represents the emotions of a player. From the LL, AIC, and BIC scores, we found that when trust  $T_{B,A}^{t-1}$ , emotions  $E_B^{t-1}$ , commitments  $C_{A,B}^{t-1}$ , and goals  $G_B^{t-1}$  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  $G_B^{t-1} \rightarrow E_B^t$  and  $C_{A,B}^{t-1} \rightarrow E_B^t$  hold. In addition, we learned two new relationships  $E_B^{t-1} \rightarrow E_B^t$  and  $T_{B,A}^{t-1} \rightarrow E_B^t$ . We obtain the AUC scores for  $E_B^t | G_B^{t-1}$  (0.69),  $E_B^t | C_{A,B}^{t-1}$  (0.5),  $E_B^t | E_B^{t-1}$  (0.56), and  $E_B^t | T_{B,A}^{t-1}$  (0.59) to find that  $G_B^{t-1}$  has the highest influence on  $E_B^t$  followed by  $E_B^{t-1}$  and  $T_{B,A}^{t-1}$ .

**Rel<sub>5</sub>**  $P(T_{B,A}^t | E_B^{t-1}, C_{A,B}^{t-1}, G_B^{t-1})$  represents trust of a player for another player. From the LL, AIC, and BIC scores, we found that when  $T_{B,A}^{t-1}$ ,  $E_B^{t-1}$ , and  $C_{A,B}^{t-1}$  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  $C_{A,B}^{t-1} \rightarrow T_{B,A}^t$  and  $E_B^{t-1} \rightarrow T_{B,A}^t$  hold and  $G_B^{t-1} \rightarrow T_{B,A}^t$  does not hold. In addition, we learned a new relationship  $T_{B,A}^{t-1} \rightarrow T_{B,A}^t$ . We obtain the AUC scores for  $T_{B,A}^t | T_{B,A}^{t-1}$  (0.74),  $T_{B,A}^t | E_B^{t-1}$  (0.63), and  $T_{B,A}^t | C_{A,B}^{t-1}$  (0.47) to find that  $T_{B,A}^{t-1}$  has the highest influence on  $T_{B,A}^t$ .

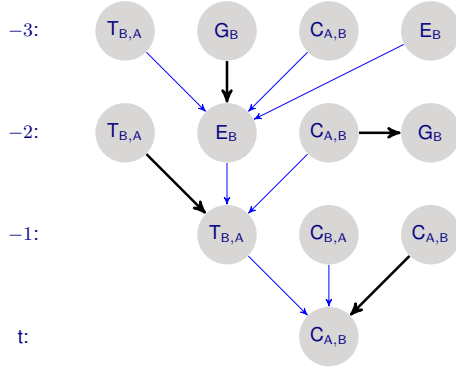
Table 4 compares the relationships in the baseline models and the new models. Figure 5 shows the new models,  $M'_1$  and  $M'_2$ , produced by combining the new relationships for the respective target variables, and highlights the relationships that yield the highest AUC scores of any relationship with the same conclusion.

## 8 Discussion and Conclusion

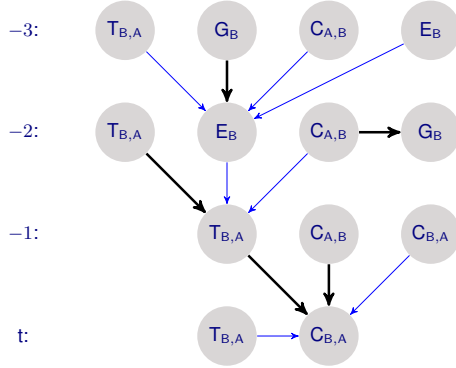
We initially construct two baseline models,  $M_1$  and  $M_2$ , based on prior work that capture the outcome of commitments based on emotions along with trust and goals. We evaluate these models using the LL, AIC, BIC, and AUC scores based on data from human subjects. Accordingly, we construct two new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

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

Baseline Relationships	Changes	New Relationships
<b>Rel<sub>1</sub></b> : $P(C_{A,B}^t   T_{B,A}^{t-1}, C_{B,A}^{t-1})$	added: $C_{A,B}^{t-1}$	$P(C_{A,B}^t   T_{B,A}^{t-1}, C_{B,A}^{t-1}, C_{A,B}^{t-1})$
<b>Rel<sub>2</sub></b> : $P(C_{B,A}^t   T_{B,A}^{t-1}, C_{A,B}^{t-1}, E_B^{t-1})$	removed: $E_B^{t-1}$ added: $C_{B,A}^{t-1}, T_{B,A}^t$	$P(C_{B,A}^t   T_{B,A}^{t-1}, C_{A,B}^{t-1}, C_{B,A}^{t-1}, T_{B,A}^t)$
<b>Rel<sub>3</sub></b> : $P(G_B^t   C_{A,B}^{t-1})$	removed: $C_{A,B}^{t-1}$ added: $C_{A,B}^t$	$P(G_B^t   C_{A,B}^t)$
<b>Rel<sub>4</sub></b> : $P(E_B^t   G_B^{t-1}, C_{A,B}^{t-1})$	added: $E_B^{t-1}, T_{B,A}^{t-1}$	$P(E_B^t   G_B^{t-1}, C_{A,B}^{t-1}, E_B^{t-1}, T_{B,A}^{t-1})$
<b>Rel<sub>5</sub></b> : $P(T_{B,A}^t   E_B^{t-1}, C_{A,B}^{t-1}, G_B^{t-1})$	removed: $G_B^{t-1}$ added: $T_{B,A}^{t-1}$	$P(T_{B,A}^t   E_B^{t-1}, C_{A,B}^{t-1}, T_{B,A}^{t-1})$



(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.

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 in the first agent influences 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 first.
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 that 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.

## 9 Acknowledgments

We thank the anonymous reviewers for their helpful comments. This research was partially supported by the US Department of Defense through the Science of Security Lablet (SoSL) at NC State University, a US Army Research Laboratory (ARL) Oak Ridge Institute for Science and Education (ORISE) Fellowship, and by IBM Research.

## References

- [Ajmeri *et al.*, 2018] N. Ajmeri, H. Guo, P.K. Murukanniah, 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*, 772–778, 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*, pp. 97–102, 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*, pp. 1031–1038, 2011.
- [Hazard and Singh, 2013] C.J. Hazard, M.P. Singh. Macau: A basis for evaluating reputation systems. *IJCAI*, pp. 191–197, 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 Schwarzenruber, 2009] E. Lorini, F. Schwarzenruber. 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.
- [Paradedda *et al.*, 2017] R. Paradedda, 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.