Bayesian Model of the Social Effects of Emotion in Decision-Making in Multiagent Systems

Celso M. de Melo
Institute for Creative Technologies, USC, 12015 Waterfront Drive, Building #4 Playa Vista, CA 90094-2536, USA
demelo@ict.usc.edu

Peter Carnevale
University of Southern California Marshall School of Business, Los Angeles, CA 90089-0808, USA
peter.carnevale@marshall.usc.edu

Stephen Read
University of Southern California Department of Psychology, Los Angeles, CA 90089-1061, USA
read@rcf.usc.edu

Dimitrios Antos
Harvard University, 33 Oxford st., Maxwell-Dworkin 217, Cambridge, MA 02138, USA
antos@fas.harvard.edu

Jonathan Gratch
Institute for Creative Technologies, USC, 12015 Waterfront Drive, Building #4 Playa Vista, CA 90094-2536, USA
gratch@ict.usc.edu

ABSTRACT

Research in the behavioral sciences suggests that emotion can serve important social functions and that, more than a simple manifestation of internal experience, emotion displays communicate one's beliefs, desires and intentions. In a recent study we have shown that, when engaged in the iterated prisoner's dilemma with agents that display emotion, people infer, from the emotion displays, how the agent is appraising the ongoing interaction (e.g., is the situation favorable to the agent? Does it blame me for the current state-of-affairs?). From these appraisals people, then, infer whether the agent is likely to cooperate in the future. In this paper we propose a Bayesian model that captures this social function of emotion. The model supports probabilistic predictions, from emotion displays, about how the counterpart is appraising the interaction which, in turn, lead to predictions about the counterpart's intentions. The model's parameters were learnt using data from the empirical study. Our evaluation indicated that considering emotion displays improved the model's ability to predict the counterpart's intentions, in particular, how likely it was to cooperate in a social dilemma. Using data from another empirical study where people made inferences about the counterpart's likelihood of cooperation in the absence of emotion displays, we also showed that the model could, from information about appraisals alone, make appropriate inferences about the counterpart's intentions. Overall, the paper suggests that appraisals are valuable for computational models of emotion interpretation. The relevance of these results for the design of multiagent systems where agents, human or not, can convey or recognize emotion is discussed.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence – *Intelligent Agents*; D.2.2 [**Software Engineering**]: Design Tools and Techniques – *User Interfaces*

General Terms

Design, Experimentation, Theory

Keywords

Emotion, Appraisals, Expression, Bayesian, Decision-Making

Appears in: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (eds.), 4–8 June 2012, Valencia, Spain.

1. INTRODUCTION

Recent developments in the behavioral sciences have led to a revolution in the understanding of the role of emotion in cognition and social behavior. Contrary to the classical view of emotion as an obstacle to rational decision-making [1, 2], this research emphasizes the positive influence emotion can have in decision-making [3-5]. As a consequence, there has been growing interest on the impact emotions can have in multiagent systems [6] and several computational models of emotion have recently been proposed [7-10]. Following the initial focus on the *intra*personal effects of emotion [11, 12], these models also focus on the impact of emotion in the self's decision-making. However, the *inter*personal effect of emotion in decision-making is also interesting and important [13-15] – i.e., the impact of another's emotions on one's decision-making. In this paper we explore a computational model for the interpersonal effect of emotion in decision-making.

A useful framework for understanding the interpersonal effect of emotion is the theory of the social functions of emotion [16-18]. This theory emphasizes that emotional expressions are not simple manifestations of internal experience; rather, expressions are other-directed and communicate one's beliefs, desires and intentions [18-21]. Emotion displays, thus, help regulate social interaction. For instance, guilt occurs when someone transgresses an accepted social norm and serves as an apology, signaling regret, which, in turn, contributes to avoid reprisals from others [22]. To study the social functions of emotion in decision-making, de Melo et al. [23, 24] conducted a series of experiments where participants engaged in a social dilemma - the iterated prisoner's dilemma [25] - with different embodied agents. Even though following the same strategy to choose their actions, the agents showed facial displays of emotion that reflected different social value orientations (e.g., cooperative or competitive). The results indicated people's decision-making was influenced by the emotion displays and people cooperated more with agents which displays reflected a desire for cooperation (e.g., smile when mutual cooperation occurred in the game) than one which displays reflected selfish desires (e.g., a smile when the agent maximized its reward at the expense of the participant). Using the empirical data collected in these studies, de Melo et al. [26] then developed, based on maximum-likelihood estimation, a computational model for decision-making in a social dilemma that took into account the outcome of the dilemma and the emotion

display. Their results showed that this model was more accurate than a model which only took into account the dilemma's outcome.

Recently, we have extended their research with two experiments that address the *mechanism* by which emotions serve their social functions. To understand this mechanism two questions needed to be answered: What is the information conveyed by emotion displays? How is this information retrieved from the displays? To answer them, we looked at appraisal theories of emotion. In appraisal theories [27], emotion displays arise from cognitive appraisal of events with respect to the agent's goals, desires and beliefs (e.g., is this event congruent with my goals? Who is responsible for this event?). According to the pattern of appraisals that occurs, different emotions are experienced and displayed. Now, since displays reflect the agent's intentions through the appraisal process, it is also plausible to ask whether people can infer from emotion displays the agent's goals by reversing the appraisal mechanism. The question then becomes: can people retrieve information about how the sender is appraising the situation from emotion displays? To address this, in our first experiment we asked participants to imagine playing the iterated prisoner's dilemma with different embodied agents. Participants were always told the same outcome occurred but were shown videos of different emotional reactions from the agent. Participants were then asked questions about how they thought the agent was appraising the situation and how likely the agent was to cooperate in the future. The results showed that participants perceived the agent to appraise the outcome consistently with expectations from appraisal theories (e.g., when the agent showed anger after an unfavorable outcome, participants perceived the agent to appraise the outcome as obstructive to its goals and to blame the participant for it). Moreover, the results showed that appraisals statistically mediated [28] the effect of emotion displays on perception of how likely the agent was to cooperate in the future. This, thus, suggests that appraisals are a key component of the information conveyed by emotion displays. To verify that perception of appraisals influence perception of the agent's likelihood of cooperation, in our second experiment we explicitly manipulated perceptions of appraisal and measured the effect on perceptions of likelihood of cooperation. The manipulation consisted of having the agents, instead of showing facial displays of emotion, express how they were appraising the outcome through text (e.g., "I really don't like this outcome and I blame you for it"). The results showed that perceptions of appraisal influenced people's perception of how likely the agent was to cooperate in the future; moreover, when the expression of appraisals corresponded, according to predictions of appraisal theories, to the emotions displayed in the first experiment, the effects on perceptions of likelihood of cooperation where very similar across experiments. Overall, these studies suggest a causal model where emotion displays lead people to infer how the agent is appraising the outcome and that, in turn, leads people to infer how likely the agent is to cooperate in the future.

In this paper we propose a computational model that captures this appraisal-based mechanism for the interpersonal effect of emotion in decision-making. The model is useful for multi-agent systems for, at least, two reasons: (1) it can be used to design agents that convey through emotion displays appropriate information about beliefs, desires and intentions; (2) it can be used by agents to interpret how the other party, human or agent, is appraising the situation and, thus, infer its intentions. At its core, the model is about *inferring*, from emotion displays, how the counterpart appraises the situation and, from this, *inferring* the other's intentions in the social encounter. Because there is a strong inductive component in this model, we follow a Bayesian approach [29]. We considered three alternative Bayesian networks: the first considered the outcome of the dilemma only; the second considered the outcome and the emotion displayed; the third considered the outcome, emotion display and appraisals. The models' parameters were learnt from the empirical data collected in the first of the aforementioned studies. We compared models with respect to their accuracy in predicting the counterpart's likelihood of cooperation in the future. Our first hypothesis, following de Melo et al.'s [26] findings was that:

Models that considered emotion display would have better accuracy than models that did not *(H1)*

However, the focus of this paper is on showing the value of integrating (perceptions of) appraisals in a model of decision-making. One important advantage appraisals provide is a structure which is shared by several emotions. For instance, *conduciveness to goals* is an appraisal which is shared by joy and sadness [27]: an event which is conducive to someone's goals causes joy; an event which is obstructive to someone's goals causes sadness. This shared structure provides a mechanism for learning parameters and making inferences regarding emotions even in the absence of examples for that particular emotion. All that is necessary is data for the emotions with which the missing emotion shares appraisals. So, our next hypothesis was:

Models that considered appraisals would have better accuracy than models that did not, over test sets which included emotions not seen in the training set *(H2)*

Finally, there are situations where people express how they are appraising a situation without resorting to emotion expression. An obvious example is when people convey verbally their attitudes toward an event. The data collected in the second of the aforementioned studies – where people convey appraisals through text – is a case in point. This dataset could, thus, be used to test our third and final hypothesis:

Models that considered appraisals could be accurate even when no emotion was shown *(H3)*

The rest of the paper is organized as follows: Section 2 presents the data in the two empirical studies; Section 3 presents the Bayesian model alternatives; Section 4 describes three experiments which test each of the hypotheses; finally, Section 5 discusses the results and its implications.

2. EMPIRICAL DATA

2.1 Study 1

The Bayesian model presented in this paper is based on data collected in a recent empirical study. In this study, we gave participants scenarios where they imagined playing the iterated prisoner's dilemma with embodied agents that displayed emotion. The payoff matrix we used is shown in Table 1. Each scenario pertained to the first round (of a 5-round game) and corresponded to a particular outcome of the game. Participants were then shown a video of how the agent reacted to the outcome. The reaction corresponded to a facial display of emotion. The agents and

emotion displays used in the experiment are shown in Figure 1. The experiment followed a mixed design with two factors: *Outcome* (between-participants) with 4 levels (one for each outcome of the game); and, *Emotion* (repeated-measures) with 5 levels (Neutral vs. Joy vs. Anger vs. Sadness vs. Guilt). In other words, each participant only saw one outcome but, was engaged with several agents that expressed different emotions. Only certain pairings of outcome and emotion where explored: (a) in mutual cooperation (CC), we considered the neutral and joy expressions; (b) when the participant was exploited (participant cooperated and agent defected, $C_H D_A$), we considered the neutral, joy and guilt expressions; (c) when the participant exploited (participant defected and agent cooperated, $D_H C_A$), we considered the neutral, anger and sadness expressions; (d) in mutual defection (DD), we considered the neutral, joy and anger expressions. Considering only a subset of the pairings allowed us to avoid unintuitive pairings (e.g., expression of anger in mutual cooperation) and reduce overall participation time.

Table 1. The prisoner's dilemma payoff matrix

		Agent	
		Cooperates	Defects
Participant	Cooperates	Agent: 5 Participant: 5	Agent: 2 Participant: 7
	Defects	Agent: 7 Participant: 2	Agent: 4 Participant: 4

For each scenario, after watching the video of the agent's reaction, participants were asked several questions about how the agent was appraising the outcome. Questions referred to three appraisal variables: a) *conduciveness to goals*, which measures whether the event is consistent or inconsistent with the individual's goals; (b) *blameworthiness*, which measures whether the self or another agent is responsible for the event; (c) *coping potential*, which measures one's ability to deal with (or control) the consequences of an event. These variables were chosen because, even though several appraisal theories have been proposed [27, 30-33], there tends to be agreement that these are critical for the emotions considered in this study: joy occurs when the event is conducive to one's goals; anger occurs when the event is not conducive to one's goals, is caused by another agent and one has power/control over it; sadness occurs when the event is not conducive to one's goals; guilt occurs when the event is not conducive to one's goals and is caused by the self. Questions were asked on a 7-point likert scale (e.g., for conduciveness to goals, 1 meant "the outcome is not conducive at all" and 7 meant "the outcome is very conducive"). Several questions were asked for each appraisal variable [30, 31, 33] but, after averaging correlated questions, only four measures remained (on a 1 to 7 scale): conduciveness to goals, participant-blameworthiness, self-blameworthiness and coping potential. Finally, before moving to the next scenario, the participant was asked one question about the agent's likelihood of cooperation in the next round (scale: 1-"not likely to cooperate at all" to 7-"very likely to cooperate"). Overall, 405 participants were recruited for this experiment, resulting in an average of 100 per outcome.

For the purposes of learning a Bayesian model, the appraisal and likelihood of cooperation questions were converted into binary format: the feature was set to 'true' if the original classification

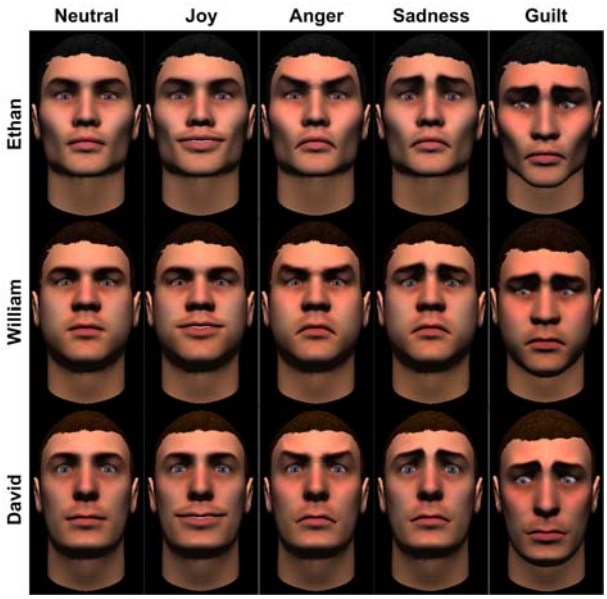

Figure 1. The facial displays of emotion.

was 5 or above; the feature was set to 'false' if the classification was 3 or below; if the classification was 4, the feature was not assigned a value (missing attribute). Each example in the training dataset, thus, had the following features:

a) Emotion Display: Neutral, Joy, Anger, Guilt or Sadness

b) Conduciveness to Goals (binary): Whether the agent was perceived to find the outcome conducive to its goals

c) Self-Blameworthiness (binary): Whether the agent was perceived to blame itself for the outcome

d) Participant-Blameworthiness (binary): Whether the agent was perceived to blame the participant for the outcome

e) Coping Potential (binary): Whether the agent was perceived to be able to deal with the consequences of the outcome

f) Likelihood of Cooperation (binary): Whether the agent was perceived to be likely to cooperate in the future

In total, excluding the examples for which the target attribute (Likelihood of Cooperation) was missing, there were 940 examples in the dataset.

2.2 Study 2

In a second empirical study, we manipulated directly how participants perceived the counterpart to be appraising the interaction, and measured perceptions of cooperation. Instead of showing emotion displays, in this study, agents expressed themselves through text in a simulated chat interface. The mapping of emotions into appraisals followed the predictions of appraisal theories [27, 30-33] and is shown in Table 2. The scenarios, game and design remained the same as in the previous study. After watching the agent's reaction, participants were asked how likely the agent was to cooperate in the next round (scale: 1-"not likely to cooperate at all" to 7-"very likely to cooperate"). Overall, 202 participants were recruited for this experiment, resulting in an average of 50 participants per outcome. The question about perception of cooperation was

discretized as in Study 1. The main difference between this and the previous dataset is that this one does not have a feature for emotion displays (or equivalently, its values are always missing). In total, the dataset had 454 examples.

Table 2. Mapping of emotion into textual expression of appraisals

Emotion	Appraisal Expression
Neutral	I neither like, nor dislike this outcome
Joy	I like this outcome
Anger	I do NOT like this outcome and I blame YOU for it
Sadness	I do NOT like this outcome
Guilt	I do NOT like this outcome and I blame MYSELF for it

3. MODELS

All Bayesian models were trained with respect to the empirical data in Study 1. Since some of the attributes in the examples could be missing (see Section 2), the EM algorithm was used for learning the parameters. The decision regarding Likelihood of Cooperation was made as follows:

- If *P(Likelihood of Cooperation) > 0.5*, true
- If *P(Likelihood of Cooperation) = 0.5*, random
- Otherwise, false

3.1 Model 1: Outcome

The first Bayesian model considered only two variables: Outcome (O) and Likelihood of Cooperation (LC). Figure 2 shows the respective Bayesian network. Outcome was set to have a uniform prior, i.e., each possible outcome occurred with 0.25 probability. The learnt parameters are shown in Table 3.

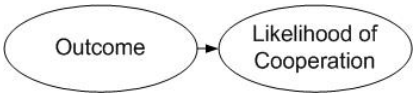

Figure 2. Bayesian network for Model 1.

Table 3. Parameters for Model 1.

O	P(LC)	O	P(LC)
CC	.470	C_HD_A	.380
DD	.405	D_HC_A	.271

3.2 Model 2: Emotion and Outcome

The second Bayesian model built on the previous and added Emotion Display (ED). Figure 3 shows the respective Bayesian network. Emotion Display was also set to have a uniform prior, i.e., each emotion occurred with 0.20 probability. The parameters are shown in Table 4.

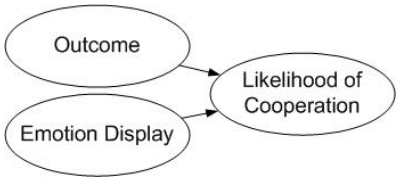

Figure 3. Bayesian network for Model 2.

Table 4. Parameters for Model 2.

ED	O	P(LC)	O	P(LC)
Neutral	CC	.235	C_HD_A	.254
Joy	CC	.719	C_HD_A	.182
Anger	CC	.500	C_HD_A	.500
Guilt	CC	.500	C_HD_A	.670
Sadness	CC	.500	C_HD_A	.500
Neutral	DD	.453	D_HC_A	.377
Joy	DD	.368	D_HC_A	.500
Anger	DD	.400	D_HC_A	.242
Guilt	DD	.500	D_HC_A	.500
Sadness	DD	.500	D_HC_A	.217

3.3 Model 3: Appraisals

The last Bayesian model added appraisal variables: Conduciveness to Goals (CG), Self-Blame (SB), Participant-Blame (PB) and Coping Potential (CP). The Bayesian network is shown in Figure 4. The appraisal variables were given BDeu priors [34], i.e., likelihood equivalent uniform Dirichlet priors. The parameters for the appraisal variables are shown in Table 5 and the parameters for Likelihood of Cooperation in Table 6.

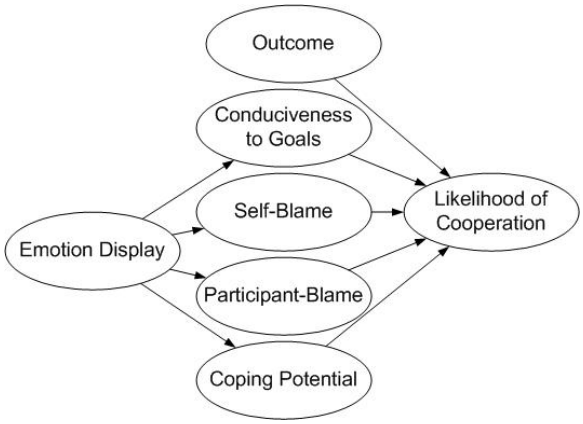

Figure 4. Bayesian network for Model 3.

Table 5. Parameters for the appraisal variables in Model 2.

ED	P(CG)	P(SB)	P(PB)	P(CP)
Neutral	.370	.203	.267	.748
Joy	.970	.206	.177	.905
Anger	.021	.381	.824	.324
Guilt	.227	.678	.222	.348
Sadness	.041	.730	.485	.285

Table 6. Likelihood of Cooperation parameters in Model 2.

CG	SB	PB	CP	O	P(LC)	O	P(LC)
T	T	T	T	CC	.436	DD	.367
F	T	T	T	CC	.082	DD	.476
T	F	T	T	CC	.410	DD	.459
F	F	T	T	CC	.129	DD	.265
T	T	F	T	CC	.837	DD	.263

F	T	F	T	CC	.002	DD	.658
T	F	F	T	CC	.640	DD	.387
F	F	F	T	CC	.324	DD	.369
T	T	T	F	CC	.146	DD	.080
F	T	T	F	CC	.259	DD	.670
T	F	T	F	CC	.054	DD	.018
F	F	T	F	CC	.172	DD	.307
T	T	F	F	CC	.990	DD	.971
F	T	F	F	CC	.014	DD	.371
T	F	F	F	CC	.776	DD	.635
F	F	F	F	CC	.203	DD	.367
T	T	T	T	C_HD_A	.320	D_HC_A	.913
F	T	T	T	C_HD_A	.849	D_HC_A	.411
T	F	T	T	C_HD_A	.084	D_HC_A	.386
F	F	T	T	C_HD_A	.528	D_HC_A	.150
T	T	F	T	C_HD_A	.108	D_HC_A	.602
F	T	F	T	C_HD_A	.863	D_HC_A	.156
T	F	F	T	C_HD_A	.243	D_HC_A	.464
F	F	F	T	C_HD_A	.526	D_HC_A	.338
T	T	T	F	C_HD_A	.502	D_HC_A	.012
F	T	T	F	C_HD_A	.366	D_HC_A	.275
T	F	T	F	C_HD_A	.335	D_HC_A	.201
F	F	T	F	C_HD_A	.383	D_HC_A	.212
T	T	F	F	C_HD_A	.642	D_HC_A	.982
F	T	F	F	C_HD_A	.821	D_HC_A	.185
T	F	F	F	C_HD_A	.122	D_HC_A	.926
F	F	F	F	C_HD_A	.398	D_HC_A	.149

4. EVALUATION

4.1 Experiment 1

To test hypothesis H1, that models which considered emotion would do better than models that did not, we tested the models accuracy with respect to the data in Study 1. Each model was re-trained using 20-fold cross-validation. The models were then compared with respect to average performance on the 20 test sets. Several standard performance measures are reported in Table 7: (a) *accuracy*, the percentage of correctly classified examples; (b) *true positives*, the number of correctly classified examples where the target (Likelihood of Cooperation) is 'true'; (c) *true negatives*, the number of correctly classified examples where the target is 'false'; (d) *false positives*, the number of incorrectly classified examples where the target is 'true'; (e) *false negatives*, the number of incorrectly classified examples where the target is 'false'. Means were compared using the 1-way independent ANOVA test.

The results showed that there was a significant difference in accuracy. In order to perform pairwise comparisons between the models, LSD post-hoc tests were performed (these are not shown in Table 7). The tests indicated that Models 2 and 3 were more accurate than Model 1. This confirmed hypothesis H1. Moreover, looking at the table, it was clear that Model 1 (based on Outcome) was making the same predictions as a game-theoretic model which always predicted defection[1]. Therefore, Outcome, by itself, seemed to be insufficient to discriminate examples in this dataset. Finally, Models 2 and 3 also seemed to be identical in their predictions. This suggested that, in this case, appraisal variables did not add more information than that provided by Emotion Display. The results also showed significant differences in the remaining variables. Looking at the true and false positive measures, it was confirmed that Model 1 always predicted defection. Still, on average, Model 1 was slightly better than Models 2 and 3, at predicting negative examples.

Table 7. Performance results for experiment 1. Means and standard deviations (in parenthesis) are shown

	acc	**tp**	**tn**	**fp**	**fn**
Model 1	62.38%	0.00	28.75	0.00	17.25
	(5.84)	(0.00)	(3.05)	(0.00)	(2.43)
Model 2	69.91%	6.15	26.05	2.70	11.10
	(7.19)	(2.08)	(3.51)	(1.66)	(3.16)
Model 3	69.91%	6.15	26.05	2.70	11.10
	(7.19)	(2.08)	(3.51)	(1.66)	(3.16)
Sig. (2-sd)	*.001**	*.000**	*.013**	*.000**	*.000**

* significant to $p<.05$

acc - accuracy; tp - true positives; tn - true negatives; fp - false positives; fn - false negatives

4.2 Experiment 2

To test hypothesis H2, that the appraisal model would have better accuracy than the others over a test set with unseen emotions, we split the data in Study 1 into two subsets: (a) the *training subset*, which included all the examples from Study 1 except the ones corresponding to Joy with the outcome C_HD_A; (b) the *test subset*, which included all the examples from Study 1 where the emotion was Joy and the outcome was C_HD_A. Models were then trained on the former subset and tested on the latter. The results are shown in Table 8.

Table 8. Performance results for experiment 2

	acc	**tp**	**tn**	**fp**	**fn**
Model 1	81.82%	0.00	72.00	0.00	16.00
Model 2	56.82%	9.00	41.00	31.00	7.00
Model 3	81.82%	0.00	72.00	0.00	16.00

acc - accuracy; tp - true positives; tn - true negatives; fp - false positives; fn - false negatives

The results showed that Model 3 was performing better than Model 2. This happened because, since there were no examples in the training set corresponding to Joy in C_HD_A, Model 2's posterior for Likelihood of Cooperation was 0.500, which corresponded to

[1] The intuition is that: the last iteration is a 1-shot prisoner's dilemma game, for which the only Nash equilibrium is mutual defection; thus, the second to last game becomes the effective last round for which a decision needs to be made. Thus, by induction, players should defect in the first round and continue doing so in every round until all rounds are completed.

a random decision. On the other hand, because of the shared appraisal structure, Model 3's posterior for Likelihood of Cooperation ($P(LC|Joy, C_HD_A)$) was 0.272. The posterior, thus, was reflecting other examples which had information about the appraisals underlying Joy. Therefore, hypothesis H2 was confirmed. Finally, the results reveal that, in this case, Model 1 performed as well as Model 3. This happened because both always defected in this test set.

4.3 Experiment 3

To test hypothesis H3, that the appraisal model could make accurate predictions even in the absence of evidence for emotion displays, we tested our models with the data from Study 2. The models were still trained on the data from Study 1 but, were tested on data from Study 2. The results are shown in Table 9.

Table 9. Performance results for experiment 3

	acc	tp	tn	fp	fn
Model 1	57.49%	0.00	261.00	0.00	193.00
Model 2	57.49%	0.00	261.00	0.00	193.00
Model 3	66.74%	72.00	231.00	30.00	121.00

acc - accuracy; tp - true positives; tn - true negatives; fp - false positives; fn - false negatives

The results showed that Model 3 was outperforming the remaining models on this dataset. This confirmed hypothesis H3. Effectively, in the absence of information about emotion displays, Model 2 could not do better than advance a prediction based only on Outcome as in Model 1.

5. DISCUSSION

This paper presents a Bayesian model that captures social effects of emotion displays in decision-making. The model's parameters were learnt using empirical data from an experiment where people engaged in a social dilemma with embodied agents that expressed emotions. The results in experiment 1 indicated that a model which took into account emotion displays was more accurate in replicating people's decision-making behavior than a model which only took into account the social dilemma outcome. This result reinforces findings in the behavioral sciences that show that non-verbal behavior – in particular, facial displays of emotion – can influence people's decision to cooperate in social dilemmas [35-38]. The results also replicate de Melo et al.'s [26] findings that a computer model of decision-making in a social dilemma improves if it takes into account the counterpart's emotion displays.

The results for Model 1, based on Outcome only and which always predicted defection, emphasize the insufficiency of a game-theoretic approach for modeling agents that interact with people. Effectively, unlike the rational prediction of defection in every round in the finite iterated prisoner's dilemma, people often cooperated in our datasets. This is compatible with the widely accepted view that people's behavior systematically deviates from game-theoretic predictions of rational behavior [39-42]. Moreover, our findings show that emotion is one of the factors that helps explain such deviations. The systematic influence of emotion displays in decision-making is, effectively, one of the premises of the social functions theory of emotion [16-18]. This theory suggests that, more than mere manifestations of internal experience, emotion expression is other-directed and communicates one's beliefs, desires and intentions. In multiagent systems research, these social effects of emotion have already been shown, for instance, when agents interact with people in social dilemmas [23, 24] and negotiation [43].

In this paper we propose further that appraisals are a useful framework to structure a computational model of emotion interpretation. Following empirical results that suggest that appraisals mediate the effect of emotion displays in decision-making, the proposed Bayesian model was structured so that variables which represented inferences about the counterpart's intentions were conditionally independent of emotion displays given information about the appraisal variables. The underlying assumption is that what matters is not the emotion display in itself but, the information it conveys about appraisals.

From a cognitive modeling perspective, it is interesting to notice that the parameters for the appraisal variables (Table 5), which represent the conditional probabilities given the emotion display, were generally in line with expectations from appraisal theories [27]: conduciveness to goals was highest for joy ($P(CG|Joy)=.970$); self-blame was highest for guilt ($P(SB|Guilt)=.678$) and sadness ($P(SB|Sadness)=.730$); participant-blame was highest for anger ($P(PB|Anger)=.824$); and, coping potential was highest for Joy ($P(CP|Joy)=.905$). This means the model was learning, from empirical data alone, some of the theoretical predictions advanced by appraisal researchers [27, 30-33].

Pragmatically, there are several advantages in following an appraisal-based model for emotion interpretation. First, appraisals provide a structure which is shared by several emotions. This provides a mechanism for learning parameters and making inferences regarding emotions even in the absence of examples for that particular emotion. The results in experiment 2 showed that the appraisal model was capable of recovering a reasonable posterior for Likelihood of Cooperation, given Joy and C_HD_A, even when no examples for that case existed in the training set. On the other hand, the model based on emotion and outcome (Model 2) could not do better than predict an even chance (0.500) of cooperation for the case where Joy is shown in C_HD_A.

A second advantage is that the appraisals model is capable of supporting inferences about the counterpart's intentions even in the absence of emotion. The results shown in experiment 3 showed that this model was capable of accurately predicting Likelihood of Cooperation for a dataset where Emotion Display was unobservable and only evidence for appraisals was available.

A third advantage of appraisals is that they provide a domain-independent mechanism for relating the counterpart's beliefs, desires and intentions to emotion displays. This relation is laid out in detail in appraisal theories of emotion [30-33] which explain how someone's beliefs, desires and intentions lead to different appraisal of situations which, in turn, lead to the experience and expression of different emotions. This knowledge can be used by multiagent system designers in, at least, two ways: (1) to implement a model, such as the one presented in this paper, that allows an agent to make inferences about the counterpart's beliefs, desires and intentions; (2) to design agents which can convey through appropriate emotions, their beliefs, desires and

intentions. Notice also that, even though appraisal theories were applied to decision-making in this paper, there is nothing in it preventing its application to other domains.

Finally, even though the paper was motivated by the literature in human-human interaction and the focus is mainly in human-agent interaction, this work has important consequences for agent-agent interaction. Simon [44] concisely articulated one of the main *intra*personal functions of emotions for intelligent agents: interrupting normal cognition when unattended goals require servicing. The theory of the social functions of emotions, on the other hand, articulates one of the main *inter*personal functions of emotions for agents: to communicate the agent's beliefs, desires and intentions. As mentioned above, appraisal theories further define how this function can be implemented through appraisals. But, why should agents use emotions to convey their mental states to other agents, as opposed to just explicitly communicate the mental states? There are many reasons, but we shall focus on two. First, from a complexity point-of-view it is more efficient for the agent to communicate information about emotions and appraisals than the whole mental state. Moreover, notice emotion need not be necessarily communicated through facial displays. Second, from an evolutionary perspective, emotion expression evolved to help solve recurrent problems that occur in social interaction [45-47]. Emotions are a quick and effective mechanism, when compared to deliberation, to respond to such problems. As multiagent systems grow in complexity, there is also an increasing need for quick and effective mechanisms to solve recurrent problems. Emotion can be one such mechanism.

6. REFERENCES

[1] Hirschman, A. 1997. The passions and the interests. Cambridge University Press.

[2] Lefford, A. 1946. The influence of emotional subject matter on logical reasoning. Journal of General Psychology 34, 127-151.

[3] Damasio, A. 1994. Descartes' error: Emotion, reason and the human brain. Putnam.

[4] Wilson, T. and Schooler, J. 1991. Thinking too much: Introspection can reduce the quality of preferences and decisions. Journal of Personality and Social Psychology 60, 181-192.

[5] Blanchette, I., Richards, A., Melnyk, L. and Lavda, A. 2007. Reasoning about emotional contents following shocking terrorist attacks: A tale of three cities. Journal of Experimental Psychology: Applied 13, 47-56.

[6] Marsella, S., Gratch, J. and Petta, P. 2010. Computational models of emotion. In A Blueprint for Affective Computing, K. Scherer, T. Banzinger and E. Roesch, Eds. Oxford University Press, Oxford, NY, 21-45.

[7] Gratch, J. and Marsella, S. 2004. A domain independent framework for modeling emotion. Journal of Cognitive Systems Research 5, 4, 269-306.

[8] Dias, J.and Paiva, A. 2005. Feeling and reasoning: A computational model for emotional agents. In Proceedings of 12th Portuguese Conference on Artificial Intelligence, EPIA 2005.

[9] Becker-Asano, C. and Wachsmuth, I. 2008. Affect simulation with primary and secondary emotions. In Proceedings of the 8th International Conference on Intelligent Virtual Agents.

[10] Wehrle, T. and Scherer, K. 2001. Toward computational modeling of appraisal theories. In Appraisal processes in emotion: Theory, methods, research, K. Scherer, A. Schorr and T. Johnstone, Eds. Oxford University Press, New York, 350-365.

[11] Loewenstein, G. and Lerner, J. 2003. The role of affect in decision making. In Handbook of Affective Sciences, R. Davidson, K. Scherer and H. Goldsmith, Eds. Oxford University Press, New York, 619-642.

[12] Blanchette, I. and Richards, A. 2010. The influence of affect on higher level cognition: A review of research on interpretation, judgment, decision making and reasoning. Cognition and Emotion 15, 1-35.

[13] Morris, M. and Keltner, D. 2000. How emotions work: An analysis of the social functions of emotional expression in negotiations. Research in Organizational Behavior 22, 1-50.

[14] Van Kleef, G., De Dreu, C., and Manstead, A. 2004. The interpersonal effects of anger and happiness in negotiations. Journal of Personality and Social Psychology 86, 57-76.

[15] Rafaeli, A. and Sutton, R. 1989. The expression of emotion in organizational life. Research in Organizational Behavior 11, 1-43.

[16] Frijda, N. and Mesquita, B. 1994. The social roles and functions of emotions. In Emotion and culture: Empirical studies of mutual influence, S. Kitayama and H. Markus, Eds. American Psychological Association, Washington, DC, 51-87.

[17] Keltner, D. and Haidt, J. 1999. Social functions of emotions at four levels of analysis. Cognition and Emotion 13, 505-521.

[18] Keltner, D. and Kring, A. 1998. Emotion, social function, and psychopathology. Review of General Psychology 2, 320-342.

[19] Bavelas, J., Black, A., Lemery C. and Mullet, J. 1986. 'I show how you feel': Motor mimicry as a communicative act. Journal of Personality and Social Psychology 50, 322-329.

[20] Fernandez-Dols, J. and Ruiz-Belda, M. 1995. Are smiles signs of happiness? Gold medal winners at the Olympic games. Journal of Personality and Social Psychology 69, 1113-1119.

[21] Kraut, R. and Johnston, R. 1979. Social and emotional messages of smiling: An ethological approach. Journal of Personality and Social Psychology 37, 1539-1533.

[22] Keltner, D. and Buswell, B. 1997. Embarrassment: Its distinct form and appeasement functions. Psychological Bulletin 122, 250-270.

[23] de Melo, C., Carnevale, P. and Gratch, J. The impact of emotion displays in embodied agents on emergence of cooperation with people. Presence: Teleoperators and Virtual Environments Journal, 2011, in press.

[24] de Melo, C., Carnevale, P. and Gratch, J. 2011. Reverse appraisal: Inferring from emotion displays who is the cooperator and the competitor in a social dilemma. In Proceedings of 33rd Annual Meeting of the Cognitive Science Society, 396-401.

[25] Poundstone, W. 1993. Prisoner's dilemma. Doubleday.

[26] de Melo, C., Carnevale, P., Antos, D. and Gratch, J. 2011. A computer model of the interpersonal effect of emotion displayed in social dilemmas. In Proceedings of the Affective Computing and Intelligent Interaction (ACII) Conference, 67-76.

[27] Ellsworth, P. and Scherer, K. 2003. Appraisal processes in emotion. In Handbook of Affective Sciences, R. Davidson, K. Scherer and H. Goldsmith, Eds. Oxford University Press, New York, 572-595.

[28] Preacher, K., & Hayes, A. 2008. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior Research Methods 40, 879-891.

[29] Griffiths, T., Kemp, C. and Tenenbaum, J. 2008. Bayesian models of cognition. In The Cambridge handbook of computational cognitive modeling, Ron Sun, Ed. Cambridge University Press.

[30] Scherer, K. 2001. Appraisal considered as a process of multi-level sequential checking. In Appraisal processes in emotion: Theory, methods, research, K. Scherer, A. Schorr and T. Johnstone, Eds. Oxford University Press, New York, 92-120.

[31] Roseman, I. 2001. A model of appraisal in the emotion system: integrating theory, research, and applications. In Appraisal processes in emotion: Theory, methods, research, K. Scherer, A. Schorr and T. Johnstone, Eds. Oxford University Press, New York, 68-91.

[32] Ortony, A., Clore, G. and Collins, A. 1988. The cognitive structure of emotions. Cambridge University Press.

[33] Smith, C. and Ellsworth, P. 1985. Patterns of cognitive appraisal in emotion. Journal of Personality and Social Psychology 48, 813-838.

[34] Heckerman, D., Geiger, D. and Chickering, D. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning 20, 197-243.

[35] Boone, R. and Buck, R. 2003. Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation. Journal of Nonverbal Behavior 27, 163-182.

[36] Frank, R. 1988. Passions within reason. Norton.

[37] Schug, J., Matsumoto, D., Horita, Y., Yamagishi, T. and Bonnet, K. 2010. Emotional expressivity as a signal of cooperation. Evolution and Human Behavior 31, 87-94.

[38] Scharlemann, J., Eckel, C., Kacelnik, A. and Wilson, R. 2001. The value of a smile: Game theory with a human face. Journal of Economic Psychology 22, 617-640.

[39] Tversky, A. and Kahneman, D. 1981. The framing of decisions and the psychology of choice. Science 211, 453-458.

[40] Simon, H. 1997. Models of bounded rationality. MIT Press.

[41] Starmer, C. 2000. Developments in non-expected utility theory: The hunt for descriptive theory of choice under risk. Journal of Economic Literature 38, 332-382.

[42] Camerer, C. 1995. Individual decision making. In Handbook of Experimental Economics, J. Kagel and A. Roth, Eds. Princeton University Press, Princeton.

[43] de Melo, C., Carnevale, P. and Gratch, J. (2011). The effect of expression of anger and happiness in computer agents on negotiations with humans. In Proceedings of Autonomous Agents and Multiagent Systems (AAMAS) 2011.

[44] Simon, H. 1967. Motivational and emotional controls of cognition. Psychological Review 74, 29-39.

[45] Darwin, C. 1872. The expression of the emotions in man and animals. Murray.

[46] Ekman, P. 1992. An argument for basic emotions. Cognition and Emotion 6, 169-200.

[47] Lazarus, R. 1991. Emotion and adaptation. Oxford University Press.