A Bayesian Model of Mind Reading from Decisions and Emotions in Social Dilemmas

Kazunori Terada (kazunori.terada@acm.org)
Gifu University, 1-1 Yanagido, Gifu 501-1193, Japan

Celso M. de Melo (celso.miguel.de.melo@gmail.com)
DEVCOM U.S. Army Research Laboratory, Playa Vista CA 90094, USA

Francisco C. Santos (Francisco.Santos@fct.pt)
INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, IST-Taguspark, 2744-016, Porto Salvo, Portugal

Jonathan Gratch (gratch@ict.usc.edu)
Institute for Creative Technologies, University of Southern California, Playa Vista CA 90094, USA

Abstract

Humans can effectively infer others' mental states, predict their behavior, and adapt their own level of cooperation accordingly in social dilemmas. However, the computational mechanisms underlying this ability remain unclear. While previous research has shown that people use both actions and emotional expressions as social cues, how these different signals are integrated during social inference has not been formally modeled. Here we propose a Bayesian framework that explains how people infer others' Social Value Orientation (SVO) from their decisions and emotional expressions in the iterated Prisoner's Dilemma. Our model formalizes this inference process through two key mechanisms: (1) rational decision-making based on utility transformation according to SVO, and (2) emotional expressions driven by outcome appraisals. We tested our model against empirical data from an existing study involving 711 participants and found that it captured both their reputation judgments and cooperation predictions. These results suggest that people may employ Bayesian inference to integrate behavioral and emotional signals when predicting others' cooperative tendencies.

Keywords: Bayesian Theory of Mind, Prisoner's Dilemma, Appraisal, Reverse Appraisal

Introduction

In social dilemmas such as the Prisoner's Dilemma, conditional cooperation—cooperating with those who also cooperate while avoiding exploitation by defectors—is theoretically optimal (Axelrod & Hamilton, 1981; Nowak, 2006). Specifically, unconditional cooperation risks exploitation by defectors, whereas unconditional defection forgoes the mutual benefits of working with trustworthy partners (Trivers, 1971; Axelrod & Hamilton, 1981). However, implementing conditional cooperation hinges on accurately predicting whether others will cooperate or defect (Fehr & Fischbacher, 2003). Because people's intentions cannot be directly observed and their past behaviors may not reliably predict future choices (Rand & Nowak, 2013), individuals rely on social cues (e.g., emotional expressions) to infer others' preferences and intentions. This makes the ability to "read minds," that is, to form accurate expectations of others' decisions, crucial for navigating social interactions successfully (Singer & Fehr, 2005; Byrne & Whiten, 1988; Tomasello, Carpenter, Call, Behne, & Moll, 2005; de Melo, Carnevale, Read, & Gratch, 2014).

Action prediction can broadly be divided into two major approaches: a *rule-based* approach and a *mentalist* approach. The rule-based approach employs a generative model of actions $p(a|s)$, learned through prior experience (Whiten, 1996; Sutton & Barto, 1998; Wood & Neal, 2007). Under this model, the observer infers the most likely action $\hat{a}$ in situation $s \in \mathcal{S}$ by

$$\hat{a} = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\ p(a|s;\phi). \tag{1}$$

Although the rule-based model fits known states, its weak inductive bias for state similarity limits generalization to novel s' and estimating $p(a|s)$ requires many (s, a) observations, implying high sample complexity.

In contrast, the mentalist approach posits that actions arise from rationally maximizing a utility function (Von Neumann & Morgenstern, 1944; Savage, 1954; Ng & Russell, 2000; Premack & Woodruff, 1978; Dennett, 1987; Gergely & Csibra, 2003; Baker, Saxe, & Tenenbaum, 2009; Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017):

$$a^* = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\ u(a, s;\psi), \tag{2}$$

where ψ represents various mental elements—beliefs, desires, goals, preferences, emotional states, and social values—that parameterize the utility function. Leveraging its goal-directed bias, the mentalist model narrows the hypothesis space and lowers sample complexity; once utility peaks are inferred, optimal actions for any novel state s' can be recomputed, yielding strong generalization. Under this view, predicting actions reduces to inferring the parameters ψ—the mental state that best explains the observed behavior in the given situation:

$$p(\psi|a, s) \propto p(a|\psi, s)p(\psi). \tag{3}$$

Observers can infer a latent mental state ψ by inverting a situation-dependent action-generation model, which yields a wide range of observable outputs, such as

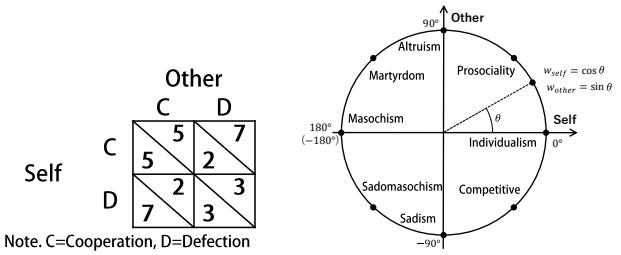
In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

bodily movements (Baker et al., 2017), decisions (Kelley & Stahelski, 1970; Van Lange & Visser, 1999; de Melo et al., 2014), and emotional expressions (Wu, Baker, Tenenbaum, & Schulz, 2017; de Melo et al., 2014). In social interactions, emotional expressions frequently serve as pivotal cues for inferring others' mental states—a process sometimes referred to as Reverse Appraisal (Gratch & de Melo, 2019; Hareli & Hess, 2010)—and thus influence the observer's subsequent behavior (van Kleef & Côté, 2021, 2018). For example, joy can foster the continuation of reciprocal relationships (Scharlemann, Eckel, Kacelnik, & Wilson, 2001; de Melo et al., 2014; de Melo & Terada, 2020), anger may prompt concessions (van Kleef, Dreu, & Manstead, 2004; Sinaceur & Tiedens, 2006; van Kleef & Côté, 2007), disappointment or sadness can elicit support or aid (Lelieveld, Van Dijk, Van Beest, & Van Kleef, 2013; Takagi & Terada, 2021), and regret can facilitate relationship repair (de Melo et al., 2014; de Melo & Terada, 2020). Recent evidence further shows that specific emotion patterns aligned with underlying social preferences can down-regulate overly cooperative behavior toward artificial agents (Ito, de Melo, Gratch, & Terada, 2024). In social dilemmas, people read joy, anger, and regret to infer intentions, update reputations (de Melo, Terada, & Santos, 2021), and adjust their own cooperation, thereby achieving conditional cooperation (de Melo et al., 2014; de Melo & Terada, 2020).

However, the computational mechanisms by which people integrate emotional expressions to infer whether a person is fundamentally cooperative or competitive—and how that affects future decisions—remain insufficiently understood. In the present study, we propose a Bayesian framework that uses both observed decisions and emotional expressions to infer a counterpart's underlying social tendencies in the Prisoner's Dilemma. Specifically, we validate this approach against data from 711 participants who interacted with an emotionally expressive counterpart in an iterated Prisoner's Dilemma. Unlike previous Bayesian Theory of Mind models, our framework explicitly treats emotions as rational signals of outcome evaluations, which arise from balancing one's own payoff against the other's payoff. As we detail in the next section, we formalize this balance through the lens of Interdependence Theory (Van Lange, 1999; Van Lange, Joireman, Parks, & Van Dijk, 2013), in particular by leveraging the concept of Social Value Orientation (SVO), which captures how individuals weigh their own payoff relative to others'. By incorporating these processes into generative models, our approach provides a concrete account of how people use actions and affective cues to make inferences in social dilemmas.

Computational Model

In non-zero-sum interdependent situations like the Prisoner's Dilemma (see Fig. 1a), SVO serves as a crucial



(a) The payoff matrix for the prisoner's dilemma

(b) The SVO ring.

Figure 1: The Prisoner's Dilemma game and SVO ring. (a) The payoff matrix showing outcomes for both players in points. (b) The SVO ring represents how individuals weigh payoffs between self (horizontal axis) and other (vertical axis). The angle θ determines the relative weights through $w_{self} = \cos\theta$ and $w_{other} = \sin\theta$. At $\theta = 0°$ (individualistic), only self-payoffs are valued; at $\theta = 45°$ (prosocial), self and other's payoffs are weighted equally.

individual characteristic ψ that influences utility. SVO represents how individuals weigh their own outcomes against others' outcomes, acting as a utility transformer that motivates cooperative or individualistic behavior in social dilemma situations (Van Lange, 1999) (see Fig. 1b). The differences in SVO lead to distinct patterns in both decision-making and emotional responses (Bogaert, Boone, & Declerck, 2008; Kramer, McClintock, & Messick, 1986; Cremer & Van Lange, 2001; Van Lange, Bekkers, Schuyt, & Van Vugt, 2007). Thus, mental state inference in the Prisoner's Dilemma is formalized as follows:

$$\begin{aligned} p(\text{SVO} \mid \text{EE}, \text{Option}, \text{Outcome}) \propto \\ p(\text{Decision} \mid \text{SVO}, \text{Option})^{\alpha} \\ p(\text{EE} \mid \text{SVO}, \text{Outcome})^{\beta} p(\text{SVO})^{\gamma}. \end{aligned} \tag{4}$$

Here, SVO is the social value orientation to be inferred. EE is the observed emotion expression. Option denotes a player's choice—C (cooperate) or D (defect); Outcome is the joint result (self, other): CC (mutual cooperation), CD (other defects), DC (self defects), or DD (mutual defection). The term $p(\text{SVO})$ represents a prior distribution over SVO angles, which encodes any initial information or biases about the counterpart's cooperative vs. competitive tendencies (i.e., their "reputation"). The exponents α, β, and γ regulate the relative influence of the likelihood terms and the prior.

The Prisoner's Dilemma involves two distinct appraisal processes: the evaluation of game options and the evaluation of game outcomes [1]. As shown in Fig. 2, these processes are represented by two likelihood

[1] Our formulation follows the two-stage appraisal structure of Ito et al. (2024).

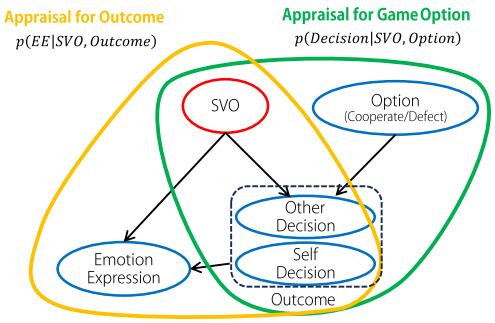


Figure 2: Generative model of the other player's decision-making and emotional processes in the Prisoner's Dilemma. The graph represents how the other player's SVO generates their decisions and emotion expressions through two appraisal processes. This generative model serves as the basis for inferring the other's SVO through Bayesian inference.

functions: $p(\text{Decision} \mid \text{SVO}, \text{Option})$ for decisions and $p(\text{EE} \mid \text{SVO}, \text{Outcome})$ for emotion expressions. For prosocials, cooperation is valued more highly than defection and mutual cooperation yields positive emotions, while for proselfs, defection is preferred and exploitation leads to positive emotions.

Once the posterior distribution over SVO angles is obtained, the probability that the counterpart will cooperate in the next round (which we interpret as their "cooperative intention") can be derived by integrating over all possible SVO angles:

$$p(\text{C}) = \int p(\text{C} \mid \text{SVO}) p(\text{SVO} \mid \text{EE}, \text{Option}, \text{Outcome}) d(\text{SVO}), \tag{5}$$

where $p(\text{C} \mid \text{SVO})$ quantifies the likelihood of rationally choosing cooperation for a given SVO.

Social Value Orientation and Utility

Let θ denote the SVO angle. We define the subjective utility $U_o(\theta)$ for each outcome $o \in CC, CD, DC, DD$ as:

$$U_o(\theta) = \cos(\theta) \cdot r_o^{self} + \sin(\theta) \cdot r_o^{other} \tag{6}$$

where r_o^{self} and r_o^{other} are the payoffs for self and other respectively. To ensure proper representation of relative preferences across all SVO angles, we introduce a softmax transformation of utilities:

$$V_o(\theta) = \frac{\exp(\kappa U_o(\theta))}{\sum_{j \in CC,CD,DC,DD} \exp(\kappa U_j(\theta))} \tag{7}$$

where κ controls the sensitivity to utility differences. This transformation ensures that utilities are always positive and sum to one, effectively capturing how different SVOs lead to different subjective evaluations of the available options.

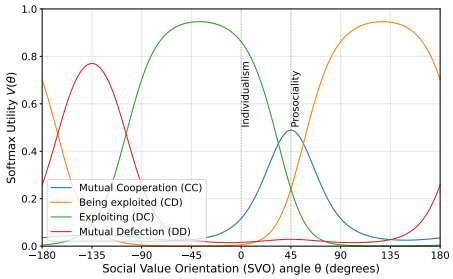


Figure 3: Softmax Utilities across SVO angles, showing how the transformation enables meaningful representation of relative preferences across all possible SVO orientations.

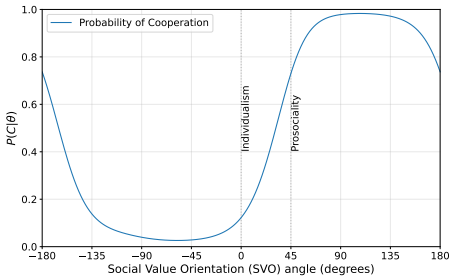


Figure 4: Probability of Cooperation $P(C|\theta)$ as a function of the SVO angle θ, showing higher cooperation probability for prosocial orientations ($\theta \approx 45°$) compared to individualistic orientations ($\theta \approx 0°$).

Decision Likelihood Function

The decision likelihood function $p(\text{Decision} \mid \text{SVO}, \text{Option})$ can be formalized using the softmax-transformed utilities. For Option = C (cooperation), the probability is denoted as $P(C|\theta)$ and defined as the sum of probabilities for outcomes involving cooperation:

$$P(C|\theta) = V_{CC}(\theta) + V_{CD}(\theta) \tag{8}$$

Similarly, for Option = D (defection), $P(D|\theta) = V_{DC}(\theta) + V_{DD}(\theta)$.

This formulation captures a key theoretical insight: the probability of cooperation emerges from aggregating the subjective values of all outcomes where cooperation is chosen, regardless of the counterpart's action. As shown in Fig. 4, this provides a formal mechanism for how social preferences translate into behavioral tendencies.

Emotion Likelihood Function

Based on these relative values, we define satisfaction $S_o(\theta)$ and dissatisfaction $D_o(\theta)$ for each outcome:

$$S_o(\theta) = \frac{V_o(\theta)}{\max(V(\theta))}, \quad D_o(\theta) = 1 - S_o(\theta). \tag{9}$$

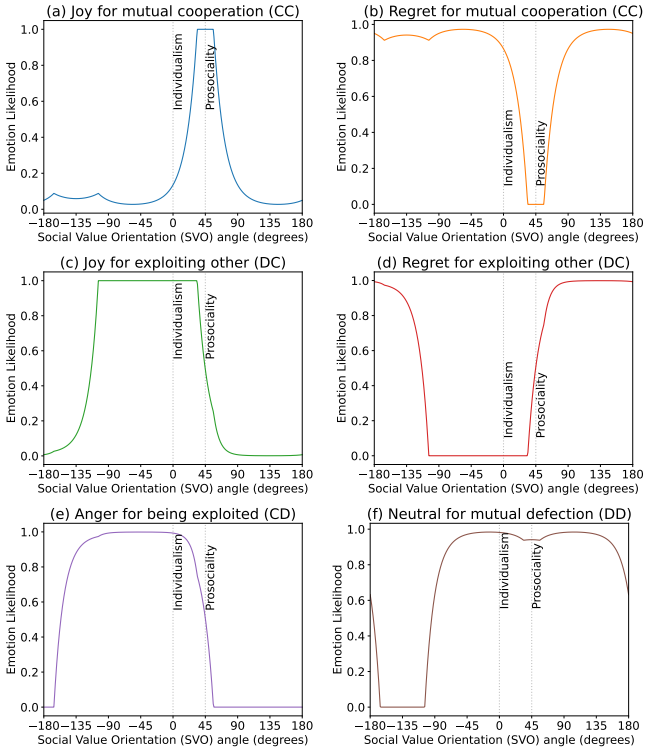


Figure 5: Emotion likelihood as a function of SVO angle for different game outcomes in the Prisoner's Dilemma. Each subplot shows how the likelihood of a specific emotion varies with SVO angle for different game outcomes.

Following appraisal theory (Scherer, Schorr, & Johnstone, 2001; Moors, Ellsworth, Scherer, & Frijda, 2013), satisfaction represents goal attainment while dissatisfaction indicates goal obstruction. For the emotion likelihood functions, we directly map these values to specific emotions:

$$\begin{aligned} P(EE = \text{Joy} \mid CC, \theta) &= S_{CC}(\theta), \\ P(EE = \text{Regret} \mid CC, \theta) &= D_{CC}(\theta), \\ P(EE = \text{Joy} \mid DC, \theta) &= S_{DC}(\theta), \\ P(EE = \text{Regret} \mid DC, \theta) &= D_{DC}(\theta), \\ P(EE = \text{Neutral} \mid DD, \theta) &= D_{DD}(\theta), \\ P(EE = \text{Anger} \mid CD, \theta) &= D_{CD}(\theta). \end{aligned} \tag{10}$$

This formulation captures key theoretical insights from appraisal theory: joy arises from goal attainment (satisfaction), regret from self-attributed goal obstruction, and anger from other-attributed goal obstruction. As shown in Fig. 5, our model predicts systematically different emotional patterns for each game outcome depending on the individual's SVO. For instance, prosocials ($\theta \approx 45°$) express *joy* after mutual cooperation (CC) (Fig. 5a), reflecting joint welfare maximization, whereas individualists ($\theta \approx 0°$) feel *regret* over missed personal gains in the same outcome (Fig. 5b). In the exploitation scenario (DC), individualists display *joy* for maximizing personal payoff (Fig. 5c), while prosocials show *regret* for undermining joint welfare (Fig. 5d). When being exploited (CD), both orientations exhibit *anger* (Fig. 5e), attributing goal obstruction to the other's action. Finally, mutual defection (DD) typically results in *neutral* (Fig. 5f) expressions due to diffused responsibility despite suboptimal outcomes. These distinct emotional patterns align with Van Kleef's EASI theory (van Kleef & Côté, 2021), indicating that emotions function as social signals that communicate how individuals evaluate outcomes and guide subsequent behaviors.

Experiment

We tested our computational model using existing data from an iterated Prisoner's Dilemma experiment where participants interacted with counterparts showing different strategies and emotional expressions. Our analysis focused on how well the model captured human inferences about others' social preferences and cooperative intentions.

Method

Behavioral Data We analyzed data from de Melo et al. (2021), which consisted of 711 participants engaging in a 20-round iterated Prisoner's Dilemma game. In their experiments, participants were randomly assigned to one of 18 conditions in a $3 \times 2 \times 3$ factorial design crossing initial reputation (negative, unknown, or positive), strategy (extortion or generosity) based on zero-determinant approaches (Press & Dyson, 2012), and emotion expression (competitive, cooperative, or neutral). Before starting the game, participants were given one of three reputational cues about their partner—negative, unknown, or positive. Based on this cue, they then rated the partner's reputation on a scale from -50 to $+50$. The counterpart's facial expressions were shown after each round according to predefined patterns aligned with either cooperative or competitive intentions, which directly corresponded to our emotion likelihood functions defined in Equation 10. The expressions for being exploited (CD: *anger*) and mutual defection (DD: *neutral*) were the same in both intention patterns. However, the expressions following mutual cooperation (CC) and exploitation (DC) differed: under cooperative intentions, the counterpart showed *joy* after CC and *regret* after DC, whereas under competitive intentions, they showed *regret* after CC and *joy* after DC.

Model-based Analysis For each participant's interaction sequence, we applied our Bayesian inference model to estimate the counterpart's SVO. The analysis involved two main steps. First, we initialized the SVO

prior using a von Mises distribution [2] centered on the initial reputation score, mapping negative reputation (-50) to individualistic orientation ($0°$), unknown reputation (0) to intermediate orientation ($22.5°$), and positive reputation ($+50$) to prosocial orientation ($45°$). We set the exponents to unity, $\alpha = \beta = \gamma = 1$, thereby giving equal weight to each likelihood term and the prior.

Second, we performed sequential updating for each round t (1 to 20). After observing the game outcome O_t (CC/CD/DC/DD) and emotion expression E_t, we updated the SVO distribution using decision likelihood $P(O_t|\theta)$ and emotion likelihood $P(E_t|O_t, \theta)$. We then computed the MAP estimate $\theta_t^* = \text{argmax}\, P(\theta|O_{1:t}, E_{1:t})$ and calculated the cooperation probability $P(C|\theta_t^*)$.

Results

Overall Results Fig. 6 shows the evolution of inferred SVO angles across 20 rounds for different combinations of strategy and emotion expression patterns, aggregated across initial reputation conditions. The density plots reveal clear patterns of SVO inference influenced by both strategy and emotion: in the Extortion–Cooperative condition, cooperative emotional expressions lead to more prosocial inferences despite the exploitative strategy, while in the Generosity–Competitive condition, competitive emotions result in more individualistic inferences despite the generous strategy. Additionally, neutral emotions produce wider variance in SVO angles compared to conditions where strategy and emotions are aligned (Extortion–Competitive or Generosity–Cooperative), which yield clearer and narrower distributions. These simulation results illustrate how incorporating emotional expressions not only shifts the inferred SVO itself but also modulates the observer's certainty about that inference.

Case Study As an example, Fig. 7 shows a single interaction where a participant started with a negative prior reputation (-50) for a generous, emotionally cooperative counterpart. Initially, the posterior distribution was centered at $0°$ (reflecting individualism). After two rounds of mutual cooperation (CC) with joy, it shifted toward prosocial. In round 3, being exploited (CD) and seeing anger partially reversed this trend, but in round 4, defection (DC) with regret restored prosocial inference. By round 20, the MAP estimate stabilized at $41°$, close to $45°$, with a cooperation probability of 0.686. This sequence illustrates how observed actions and emotions jointly shape inferences about the counterpart's SVO.

Model Validation To validate our computational model's predictions against human judgments, we conducted two analyses. First, we examined how well the model's inferred SVO angles corresponded to participants' final reputation ratings of their counterparts. Second, we compared the model's predictions of cooperation probability with participants' actual expectations of cooperation in the final round. Fig. 8 shows these comparisons.

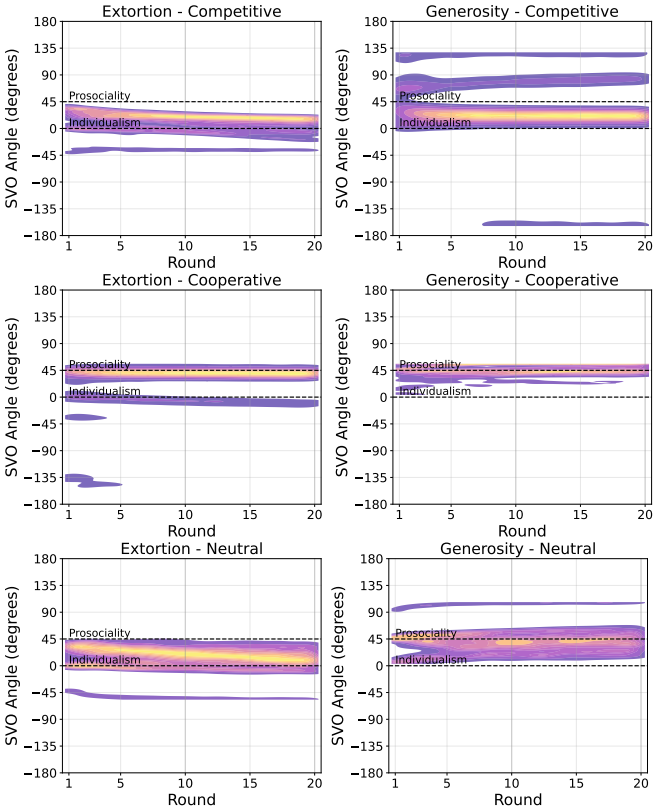


Figure 6: Evolution of inferred SVO angles across experimental conditions. Each subplot corresponds to a combination of counterpart strategy (Extortion or Generosity) and emotion (Competitive, Cooperative, or Neutral). Data are aggregated (averaged) over all initial reputation conditions. Heat maps illustrate the density of SVO estimates, with brighter colors representing higher density.

Fig. 8a reveals a significant positive correlation between model-inferred SVO angles and human reputation ratings ($r = 0.36$, $p < 0.01$). In the study by de Melo et al. (2021), participants re-evaluated the counterpart's reputation after 20 rounds, and they reported that these final ratings varied systematically according to the counterpart's strategy (extortion or generosity) and emotional expressions. The close alignment between our model's SVO estimates and the observed final reputation scores suggests that participants integrated observed actions and emotional cues in a manner consistent with Bayesian inference when updating their judgments.

Fig. 8b shows the relationship between the model's predicted probability of cooperation in the final round and participants' final-round expectations of coopera-

[2]The von Mises distribution is the circular analog of the normal distribution and serves as a close approximation to the wrapped normal distribution.

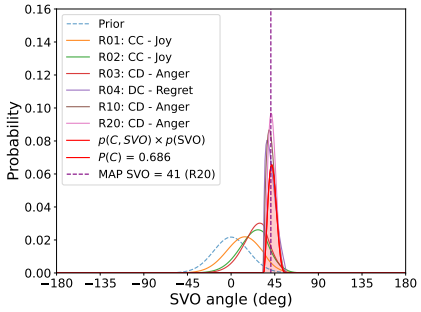
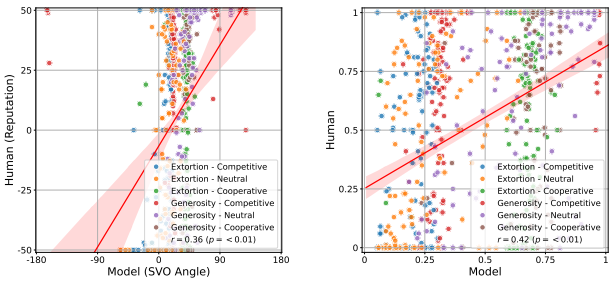


Figure 7: Evolution of SVO inference for a single participant. The plot shows posterior distributions at key rounds, the final MAP estimate (purple dashed line), and the cooperation probability calculation (red curve and shading). Prior distribution (blue dashed line) reflects initial negative reputation.



(a) Correlation between model-inferred SVO angles and human reputation ratings.

(b) Correlation between model-predicted cooperation probability and human expectations.

Figure 8: Model validation against human judgments. Colors indicate different combinations of strategy and emotion expression patterns.

tion ($r = 0.42$, $p < 0.01$). This moderate correlation indicates that the model partly captures how humans integrate strategy and emotion information to anticipate others' cooperative intentions. Notably, the distribution of points shows clear clustering by strategy, with generosity conditions yielding higher cooperation predictions from both the model and humans compared to extortion conditions.

Discussion

Our study provides a computational account of how humans integrate behavioral and emotional signals to infer others' social preferences in the iterated Prisoner's Dilemma. By developing a Bayesian framework that combines two key processes - rational decision-making based on social preferences and emotional expressions based on outcome appraisals. Specifically, our model explains the cognitive mechanisms by which people update their beliefs about others' SVO based on observed actions and emotional expressions, and demonstrates how these inferences systematically shape predictions about future cooperation.

Our model validation analysis yielded two main results. First, we found a moderate correlation between our model's inferred SVO angles and participants' reputation ratings after the game ($r = 0.36$). This suggests that our computational framework partly captures how people form impressions of others' social preferences through observed behaviors and emotional expressions during the game. Second, the model's predictions about cooperation probability, derived from rational action selection based on inferred social preferences (Equation 2), showed a moderate correlation with participants' reported expectations ($r = 0.42$). This alignment suggests that people may predict others' behavior by inferring their underlying preferences and assuming rational decision-making—a core principle of our Bayesian framework. The observation that this relationship persisted across different experimental conditions suggests some robustness in this cognitive mechanism.

Our work contributes to the broader understanding of human cooperation. While various mechanisms have been proposed to explain why self-interested individuals engage in cooperative behavior—including indirect reciprocity and reputation systems (Nowak, 2006; Alexander, 1987)—our model adds a cognitive computational perspective to this literature. Specifically, we provide a computational foundation for previous findings on the social effects of emotion expressions in strategic interaction (de Melo et al., 2014; de Melo & Terada, 2020; de Melo et al., 2021). While these studies demonstrated empirically that emotions influence cooperation, our model formalizes the cognitive mechanisms underlying this influence through Bayesian inference. Specifically, we provide a mathematical framework that captures how emotions serve as informative signals about others' appraisals of social outcomes (van Kleef & Côté, 2021), enabling observers to infer others' underlying social preferences and predict their future behavior.

Our study has several limitations that may help explain the moderate correlations observed in our validation analysis. First, we did not account for individual differences and uncertainty in social inference; for instance, some individuals may prioritize actions over emotions, disregard initial reputation, or misunderstand social cues. Second, although we theoretically derived the likelihood functions, we did not empirically validate them (e.g., with self-reports or physiological measures). Finally, our investigation was limited to three specific emotions (joy, regret, anger) within a particular Prisoner's Dilemma context, leaving other potentially relevant emotions (e.g., sadness, disappointment, guilt) and other interaction scenarios unaddressed. Addressing these limitations may yield a more comprehensive and generalizable model of how people integrate social signals to infer others' intentions in strategic interactions.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP24H00718 and JST CREST Grant Number JPMJCR21D4, Japan, and the Army Research Office under Cooperative Agreement Number W911NF-25-2-0040. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Alexander, R. D. (1987). *The biology of moral systems*. Aldine De Gruyter.

Axelrod, R., & Hamilton, W. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396. doi: 10.1126/science.7466396

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017, mar). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064. doi: 10.1038/s41562-017-0064

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009, dec). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. doi: 10.1016/j.cognition.2009.07.005

Bogaert, S., Boone, C., & Declerck, C. (2008, sep). Social value orientation and cooperation in social dilemmas: A review and conceptual model. *British Journal of Social Psychology*, *47*(3), 453–480. doi: 10.1348/014466607x244970

Byrne, R. W., & Whiten, A. (Eds.). (1988). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford Science Publications.

Cremer, D. D., & Van Lange, P. A. M. (2001). Why prosocials exhibit greater cooperation than proselfs: the roles of social responsibility and reciprocity. *European Journal of Personality*, *15*(S1), S5–S18. doi: 10.1002/per.418

de Melo, C. M., & Terada, K. (2020, sep). The interplay of emotion expressions and strategy in promoting cooperation in the iterated prisoner's dilemma. *Scientific Reports*, *10*, 1–8. doi: 10.1038/s41598-020-71919-6

de Melo, C. M., Terada, K., & Santos, F. C. (2021). Emotion expressions shape human social norms and reputations. *iScience*, *24*, 1–9. doi: 10.1016/j.isci.2021.102141

de Melo, C. M., Carnevale, P. J., Read, S. J., & Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. *Journal of Personality and Social Psychology*, *106*(1), 73–88. doi: 10.1037/a0034251

Dennett, D. C. (1987). *The intentional stance*. MIT Press.

Fehr, E., & Fischbacher, U. (2003, oct). The nature of human altruism. *Nature*, *425*(6960), 785–791. doi: 10.1038/nature02043

Gergely, G., & Csibra, G. (2003, Jul). Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292. doi: 10.1016/S1364-6613(03)00128-1

Gratch, J., & de Melo, C. M. (2019). Inferring intentions from emotion expressions in social decision making. In *The social nature of emotion expression* (pp. 141–160). Springer International Publishing. doi: 10.1007/978-3-030-32968-6_8

Hareli, S., & Hess, U. (2010, jan). What emotional reactions can tell us about the nature of others: An appraisal perspective on person perception. *Cognition & Emotion*, *24*(1), 128–140. doi: 10.1080/02699930802613828

Ito, R., de Melo, C. M., Gratch, J., & Terada, K. (2024). Emotional expression help regulate the appropriate level of cooperation with agents. In *The 12th international conference on affective computing and intelligent interaction (ACII '24)*. doi: 10.1109/ACII63134.2024.00008

Kelley, H. H., & Stahelski, A. J. (1970). Social interaction basis of cooperators' and competitors' beliefs about others. *Journal of Personality and Social Psychology*, *16*(1), 66–91. doi: 10.1037/h0029849

Kramer, R. M., McClintock, C. G., & Messick, D. M. (1986, sep). Social values and cooperative response to a simulated resource conservation crisis. *Journal of Personality*, *54*(3), 576–582. doi: 10.1111/j.1467-6494.1986.tb00413.x

Lelieveld, G.-J., Van Dijk, E., Van Beest, I., & Van Kleef, G. A. (2013). Does communicating disappointment in negotiations help or hurt? solving an apparent inconsistency in the social-functional approach to emotions. *Journal of Personality and Social Psychology*, *105*(4), 605–620. doi: 10.1037/a0033345

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013, mar). Appraisal theories of emotion: State of the art and future development. *Emotion Review*, *5*(2), 119–124. doi: 10.1177/1754073912468165

Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the seventeenth international conference on machine learning* (p. 663–670). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Nowak, M. A. (2006, dec). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560–1563. doi: 10.1126/science.1133755

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences*, *4*, 515-526. doi:

10.1017/S0140525X00076512

Press, W. H., & Dyson, F. J. (2012, may). Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences*, *109*(26), 10409–10413. doi: 10.1073/pnas.1206569109

Rand, D. G., & Nowak, M. A. (2013, aug). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413–425. doi: 10.1016/j.tics.2013.06.003

Savage, L. (1954). *The foundations of statistics*. Wiley Publications in Statistics.

Scharlemann, J. P., Eckel, C. C., Kacelnik, A., & Wilson, R. K. (2001). The value of a smile: Game theory with a human face. *Journal of Economic Psychology*, *22*(5), 617 - 640. doi: 10.1016/S0167-4870(01)00059-9

Scherer, K. R., Schorr, A., & Johnstone, T. (Eds.). (2001). *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.

Sinaceur, M., & Tiedens, L. Z. (2006). Get mad and get more than even: When and why anger expression is effective in negotiations. *Journal of Experimental Social Psychology*, *42*(3), 314–322. doi: 10.1016/j.jesp.2005.05.002

Singer, T., & Fehr, E. (2005, April). The neuroeconomics of mind reading and empathy. *American Economic Review*, *95*(2), 340–345. doi: 10.1257/000282805774670103

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.

Takagi, H., & Terada, K. (2021). The effect of anime character's facial expressions and eye blinking on donation behavior. *Scientific Reports*, *11*, 1–8. doi: 10.1038/s41598-021-87827-2

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *BEHAVIORAL AND BRAIN SCIENCES*, *28*, 675?735. doi: 10.1017/S0140525X05000129

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57.

Van Lange, P. A. M. (1999, aug). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology*, *77*(2), 337–349. doi: 10.1037/0022-3514.77.2.337

Van Lange, P. A. M., Bekkers, R., Schuyt, T. N. M., & Van Vugt, M. (2007, nov). From games to giving: Social value orientation predicts donations to noble causes. *Basic and Applied Social Psychology*, *29*(4), 375–384. doi: 10.1080/01973530701665223

Van Lange, P. A. M., Joireman, J., Parks, C. D., & Van Dijk, E. (2013, mar). The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*, *120*(2), 125–141. doi: 10.1016/j.obhdp.2012.11.003

Van Lange, P. A. M., & Visser, K. (1999, oct). Locomotion in social dilemmas: How people adapt to cooperative, tit-for-tat, and noncooperative partners. *Journal of Personality and Social Psychology*, *77*(4), 762–773. doi: 10.1037/0022-3514.77.4.762

van Kleef, G. A., & Côté, S. (2007). Expressing anger in conflict: When it helps and when it hurts. *Journal of Applied Psychology*, *92*(6), 1557–1569. doi: 10.1037/0021-9010.92.6.1557

van Kleef, G. A., & Côté, S. (2018, jan). Emotional dynamics in conflict and negotiation: Individual, dyadic, and group processes. *Annual Review of Organizational Psychology and Organizational Behavior*, *5*(1), 437–464. doi: 10.1146/annurev-orgpsych-032117-104714

van Kleef, G. A., & Côté, S. (2021, jul). The social effects of emotions. *Annual Review of Psychology*, *73*(1). doi: 10.1146/annurev-psych-020821-010855

van Kleef, G. A., Dreu, C. K. W. D., & Manstead, A. S. R. (2004). The interpersonal effects of anger and happiness in negotiations. *Journal of Personality and Social Psychology*, *86*(1), 57-76. doi: 10.1037/0022-3514.86.1.57

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

Whiten, A. (1996). When does smart behaviour-reading become mind-reading? In P. Carruthers & P. K. Smith (Eds.), *Theories of theories of mind* (pp. 277–292). Cambridge University Press. doi: 10.1017/CBO9780511597985.018

Wood, W., & Neal, D. T. (2007). A new look at habits and the habit-goal interface. *Psychological Review*, *114*(4), 843–863. doi: 10.1037/0033-295x.114.4.843

Wu, Y., Baker, C. L., Tenenbaum, J. B., & Schulz, L. E. (2017, oct). Rational inference of beliefs and desires from emotional expressions. *Cognitive Science*. doi: 10.1111/cogs.12548