A Generative Model of People's Intuitive Theory of Emotions

Sean Dae Houlihan (daeda@mit.edu), Max Kleiman-Weiner, Joshua B. Tenenbaum & Rebecca Saxe
MIT Department of Brain and Cognitive Sciences,
77 Massachusetts Ave., Cambridge, MA 02139 USA

Abstract

We present a formal model of emotion predictions that effectively captures observers' nuanced intuitions about the emotional experiences of players in a strategic and socially charged situation—a high-stakes public one-shot prisoner's dilemma. We first incorporate social inequity concerns, reputational considerations, and monetary utility in an inverse planning framework to model observers' intuitions about the latent motivational structure of gameplay. Next, in a novel approach to inverse planning models, simulated agents react to the game's outcomes, generating prediction errors. These reactions, achieved utilities, and counterfactuals are then translated into forward predictions about players' emotions. The emotion predictions generated by the model reflect the richly structured reasoning that observers exhibit when considering players' experiences, including the counterfactual dependencies of Relief and Regret, the social cognitive dependence of Envy, the prosocial dimension of Joy, and the moral content of Guilt and Embarrassment. The model captures these intuitions using a psychologically plausible architecture that resembles observers' direct judgments of the latent parameters.

Keywords: emotion; appraisal; social; inverse planning

Thinking About Emotions in Rich Social Games

Imagine watching two contestants on a game show, Arthur and Belle, play a high-stakes public one-shot social game called 'Split or Steal'. On the table is a pot of $100k. Each player can secretly choose to Split (cooperate) or Steal (defect) (C or D, respectively). If both players choose to Split, each takes home $50k. If both choose to Steal, they both leave with nothing. But if one chooses to Split and the other chooses to Steal, the one who stole takes the entire $100k and the other player leaves with nothing. Arthur and Belle both publicly promise to Split, making grand appeals to their rectitude, in a bid to convince each other to not make the game-theoretic best response. But such talk is cheap. What will they really choose? What do you (the observer) expect Arthur (the player) to feel if they both cooperate? Typical answers include joy, surprise, and relief (figure 1a). By contrast, if Belle split but Arthur stole, observers predict he will feel less joy and more guilt. The aim of this work is to formally model how human observers generate these structured emotion predictions.

Consistent with prior work, we propose that observers use an intuitive theory of mind to predict others' emotions. Ong et al. (2015) showed that observers can infer what another person wants (e.g. money) and expects (e.g. the subjective probability of winning) and that these are key features of the intuitive theory of emotion for simple lottery events. For instance, given payoffs ranging from $25 to $100 with known probabilities, observers predicted that a player would feel happiest when he (a) won more money (actual reward), and (b) won more money than he expected (inferred reward prediction error). A formal model incorporating actual reward and prediction error captured 75% of the variance in observers' predictions about players' emotions.

The Ong et al. model was ground-breaking, yet has serious limitations which restricts the space of emotions that human observers readily attribute. Lottery players make no choices, cannot be harmed, and have no social interactions. Happiness and disappointment are the main emotions predicted in response to a small sums lottery; lotteries cannot reveal observers' intuitive theory of social emotions like guilt and envy.

Model Structure: Selection of Base Features

Humans optimize for more than personal financial gain, even in one-shot and anonymous social interactions (and critically, observers recognize this). For example, the 'Split or Steal' game described above creates complex emotions in Arthur because his motives as a player are not solely to maximize his payoff in the game. Arthur may additionally want (i) for Belle to also have a good outcome (prosocial preferences), and (ii) not to be betrayed by Belle and left with the sucker's payoff. We propose that these same values are incorporated in observers' predictions about emotions. For example, when the players' choices are revealed, observers predict Arthur will feel joy not only because he won money, but also because Belle won money (figure 1a).

Fehr and colleagues have proposed that humans are motivated, to various degrees, by two kinds of concerns for fairness in social interactions (Fehr & Schmidt, 1999). Disadvantageous inequity aversion (DIA), a preference not to end up worse off than others, is a powerful and culturally universal social preference. In addition, people's choices reflect advantageous inequity aversion (AIA), a preference not to extract more than one's fair share of a resource. Here, we use Fehr's parameters of real choices as the basis for the intuitive theory of others' choices. That is, we assume that people have an intuitive grasp of the structure of people's social motives.

Generative Model of Observer's Intuitions About One-Shot Anonymous Play

In the `Base Model`, we simulate gameplay based on observers' intuitions about players' valuation of personal monetary reward and social inequity. Agents are generated by sampling a vector of preference weights ($\vec{\omega}_b$) that define how

much agents value each base feature ($Money$, AIA, DIA) during decision-making. Each base weight is sampled from a prior belief about players' preferences: $\omega_{i,\mathrm{b}} \sim \mathrm{Beta}(\alpha_{i,\mathrm{b}}, \beta_{i,\mathrm{b}})$.

Agents additionally sample a subjective belief about how likely the opponent is to defect, given by $\pi_{a_2} = P(a_2 = D)$ where a_2 is the opponent's action. An agent calculates the expected utility associated with each action (C or D) based on its preferences, its estimated probability of the other player's action, and the monetary payoff to the agent for expected outcome (payoffs are logarithmically transformed to reflect people's non-linear valuation functions). The expected utility for a given decision is thus:

$$\mathbb{E}[U_{\mathrm{b}}(a_1)] = \sum_{a_2 \in \{C,D\}} \pi_{a_2} \cdot \Big(\omega_{Money,\mathrm{b}} \cdot \ln\Big(\theta_{Money} \cdot pot + 1\Big) + \omega_{AIA,\mathrm{b}} \cdot \ln\Big(\theta_{AIA} \cdot pot + 1\Big) + \omega_{DIA,\mathrm{b}} \cdot \ln\Big(\theta_{DIA} \cdot pot + 1\Big)\Big) \tag{1}$$

where a_1 is the agent's action and θ (how the outcome loads into the base feature), is a function of a_1 and a_2. Simulated decisions follow probabilistically as samples from the softmax distribution of the expected utility: $P(a_1 \mid \vec{\omega}_b, \pi_{a2}) \propto \exp(\lambda \cdot \mathbb{E}[U_b(a_1)])$, where the optimally parameter λ is fixed at 2.

Inferring Latent Motives Through Model Inversion

Our model uses inverse planning to infer a player's values and expectations from his choice (before the opponent's choice is known). As observers only see the player's choice, to infer the player's values observers must solve an ill-posed inductive problem that involves reasoning backwards from sparse data (e.g. choosing to cooperate) to rich representations (Baker et al., 2017). We propose that observers systematically infer players' values and beliefs from their choices by inverse planning, and that these inferences are the basis for observers' predictions about players' emotions. We presented Amazon mTurk volunteers with one player's choice and the pot size, then asked them to directly rate the player's preferences and the player's estimate of the opponent's action. Observers' ratings closely matched the values obtained by inverting the generative model of play according to Bayes' rule, $P(\vec{\omega}_b, \pi_{a_2} \mid a_1) \propto P(a_1 | \vec{\omega}_b, \pi_{a_2}) \cdot P(\vec{\omega}_b, \pi_{a_2})$.

Extending Decisions to Public Games

Thus far, the `Base Model` has aimed to capture observers' intuitions about players' first-order utility preferences for monetary and social equity outcomes of the game. However, we hypothesize that observers intuit that players also have second-order preferences for how they would like to be perceived by others. For example, Arthur may choose to cooperate primarily to signal his cooperativeness to future social partners. To incorporate reputation concerns we follow a cognitively natural strategy similar to Kleiman-Weiner et al. (2017), whereby we model reputation as people's desire to be perceived in a positive light. In order for Belle to choose an action that is reputation enhancing, she must first infer how that action will be perceived by others. This requires an embedded inference loop. The inferences an observer would make about the weights of a player's base utility function are themselves weighted and treated as "second-order" utilities: a preference for certain inferences that observers make about their values. Agents now calculate expected utilities for each available action by incorporating these reputational utilities according to a vector of individually weighted preferences, $\vec{\omega}_r$.

$$\mathbb{E}[U_{\mathrm{b+r}}(a_1)] = \mathbb{E}[U_{\mathrm{b}}(a_1)] + \Big(-\omega_{Money,\mathrm{r}} \cdot \mathbb{E}(\omega_{Money,\mathrm{b}} | a_1) + \omega_{AIA,\mathrm{r}} \cdot \mathbb{E}(\omega_{AIA,\mathrm{b}} | a_1) + \omega_{DIA,\mathrm{r}} \cdot \mathbb{E}(\omega_{DIA,\mathrm{b}} | a_1) \Big) \cdot \ln(pot + 1) \tag{2}$$

This `Base+Reputation Model`[1] is obviously much richer than is necessary to predict players' choices in a Prisoner's Dilemma (which can be captured by extremely simple models with one parameter), but we propose that this richness is necessary to capture the predictions that observers make about players' emotions.

Calculating Prediction Errors via Inverse Planning

We generate a feature space that directly supports emotion predictions. When Arthur decides to cooperate, his expected utility for disadvantageous inequity must include the possibility that Belle will defect. Once Belle's choice to cooperate is revealed, Arthur has a positive prediction error: less disadvantageous inequity that he expected. This prediction error is a feature that can be used to predict his emotions. By analogy to inverse planning (inferring beliefs and desires from actions), we call this step "inverse appraisal" (inferring the effect of an event on a person's goals, beliefs, costs, and norms). The appraisal feature space includes the subjective utility of the reputation enhancement or cost on abstract values like AIA and DIA, and counterfactuals on each player's actions. We suggest that this appraisal space offers a useful level of abstraction as the features are readily expressed in natural language and are introspectable to observers (Scherer & Meuleman, 2013), thus bridging formalization and intuition.

Generative Model of Forward Emotion Predictions

In the final step we asked human participants to predict player's emotions across a range of events, pot sizes, actions, and opponent choices. We sought to qualitatively capture those emotion predictions by glossing emotion labels in terms of the features generated by the `Inverse Appraisal Model`.

[1] The sign on the terms denote the desirability of the reputation. Similar to the `Base Model`, preference weights are sampled from Beta priors, decisions follow probabilistically from the softmax distribution over decisions, and model inversion gives the conditional joint distribution of agents' preferences and beliefs given the observed decision.

We presented Amazon mTurk volunteers (n=132) with descriptions of public one-shot 'Split or Steal' games, including both players' choices and payoffs spanning five orders of magnitude ($0 – $200k). Volunteers predicted the emotional reactions of 12 players from different games by rating the intensity of 20 emotions on a continuous scale (eight of these emotions shown in figure 1a).

Crucially, observers' emotion predictions cannot not be captured by a model limited to selfish monetary payoffs and prediction errors alone, e.g. observers predicted that players who defect will feel *guilt* regardless of their financial outcome and only expect players to feel *envy* when the their outcome is worse than their opponents'. We glossed forward emotion predictions as functions of a small number of hand-coded inverse appraisal features, pulling heavily from prior work in appraisal theory (Scherer & Meuleman, 2013). No fitting procedure was used. Rather, we only assessed whether simple additive combinations of the input features with no free parameters captures relationships between the empirical emotion predictions[2].

For example, *Joy* is defined as the sum of the subjective utilities on $Money$ and AIA, and the prediction error on $Money$, which can be expressed in natural language (from a player's first person perspective) as "I win money, I get more money than expected, I don't exploit the other person". *Envy* is the negative subjective utility on DIA, or "my outcome is inferior to the other person's." *Fury* is the negative prediction errors on $Money$ and DIA, or "I get less money and am more inferior than I expected." *Regret* is the monetary counterfactual on my choice and my opponent's choice, and the negative prediction error on DIA, or "I should have made the other choice, I wish that the other player made the other choice, I am more inferior than I expected."

These generative predictions illustrate the rich expressive capacity of this feature space. For example, *fury* is increased by negative monetary prediction error, but differs from *disappointment* (negative prediction error on $Money$) in that it also increases when the player is more inferior than expected. *Joy* is increased by monetary reward and prediction error, as with Ong et al., but is decreased by exploiting the other player. This simple additive combination of features (with no fitting or weighting) shows promising similarities with human emotion predictions. We propose that the `Inverse Appraisal` process computationally recapitulates cognitive mechanisms that humans use when predicting others' complex emotions from events, actions, and outcomes.

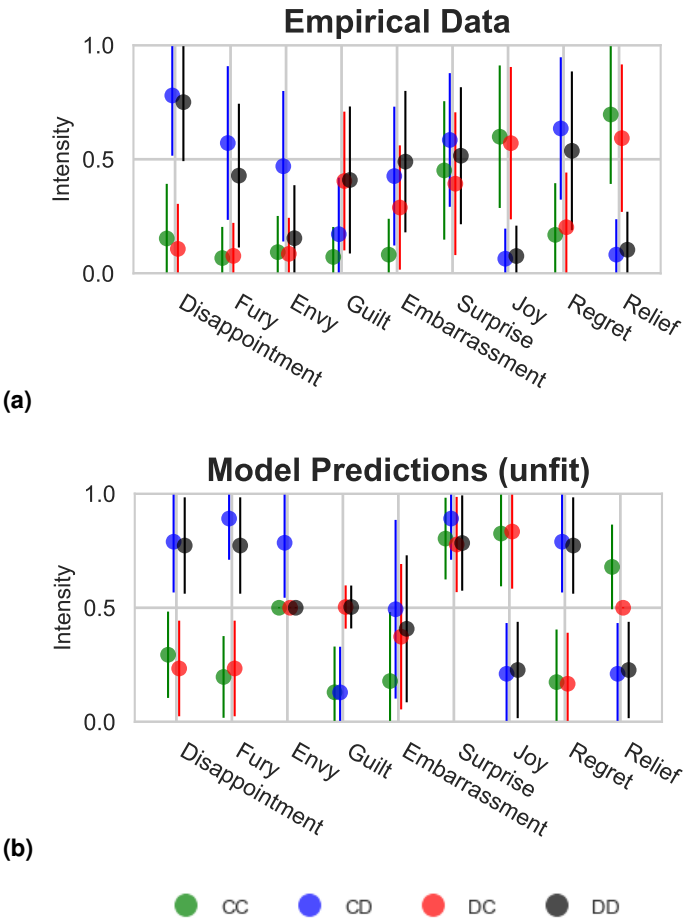

Figure 1: Emotion predictions. (a) Human observers' predictions given the game outcome. (b) The model's predictions capture key relationships in the predicted intensity within each emotion. Emotions defined as additive combinations of hand-coded input with no fitting. The pot size has been marginalized out of both the human and model data. Legend gives the judged player's action followed by the opponent's action (C: cooperate, D: defect), e.g. DC indicates the judged player defected (stole) and the opponent cooperated (split).

References

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

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, *114*(3), 817–868.

Kleiman-Weiner, M., Shaw, A., & Tenenbaum, J. B. (2017). Constructing social preferences from anticipated judgments: When impartial inequity is fair and why? In *Proceedings of the 39th annual conference of the cognitive science society.*

Ong, D. C., Zaki, J., & Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition*, *143*, 141–162.

Scherer, K. R., & Meuleman, B. (2013). Human emotion experiences can be predicted on theoretical grounds: evidence from verbal labeling. *PloS one*, *8*(3), e58166.

[2]Model predictions are transformed into the [0,1] range of the empirical judgments with a logistic function.