Intervening on Emotions by Planning Over a Theory of Mind

Tony Chen[*1], Sean Dae Houlihan[*3,1], Kartik Chandra[2], Joshua B. Tenenbaum[1], Rebecca Saxe[1]

{thc, daeda, kach, jbt, saxe} @ mit.edu

[1] MIT Brain and Cognitive Sciences, [2] MIT CSAIL, [3] Dartmouth College, Psychological and Brain Sciences

[*] denotes equal contribution

Abstract

Much of social cognition involves reasoning about others' minds: predicting their reactions, inferring their feelings, and explaining their behavior. By representing mental contents like beliefs, desires, and emotions, Bayesian Theory of Mind models have made progress in capturing how humans manage these cognitive feats. But social life is not merely observation: humans must also plan to intervene on these same mental contents. The present work models how people choose interventions to influence others' emotions. Building on a prior model of people's intuitive theory of emotions, we model how people use their intuitive theory to evaluate and simulate the effects of different interventions. We apply our model to data from behavioral experiments requiring counterfactual and joint interventions, and show a close alignment with human choices. Our results provide a step towards a potentially unifying explanation for emotion prediction and intervention, suggesting that they could arise from the same underlying generative model.

Keywords: theory of mind; emotion; social cognition; decision making and planning

Introduction

Suppose that your younger sibling just broke their favorite toy and you want to make them feel better. You carefully weigh your options. You could get them a new version of the same toy. Or, you could give them something of yours that they have shown interest in. Which one do you pick?

Moments like these remind us of a key feature of human decision making and planning: many of the actions that we take in our daily life are designed to make someone feel something specific. For our actions to be effective, we need to consider the intricacies of other people's beliefs, desires, plans, and emotions. To think about, and interact with, others, we are in essence using a mental model of their minds. Our mental models of other minds are causal, integrating core intuitions about factors such as effort, costs, constraints, and rewards, and also flexible, able to be adapted to accommodate idiosyncrasies of persons and situations.

Specifically, much of social cognition is oriented around changing and regulating others' emotions. We plan actions with goals of cheering up, calming down, and comforting those around us (Thoits, 1996; Tran et al., 2023). We may attempt to induce feelings of guilt or shame as a way of punishing others (Nelissen & Zeelenberg, 2009), praise someone as a way of rewarding them (Wu et al., 2021), or portray a tragic character to make an audience cry (Chandra et al., 2023). These everyday social goals require a great deal of cognitive sophistication. People need to consider the affordances in the environment, reason over various possible world states and how they might be reached, imagine why someone might act one way or another, and predict how others would emotionally react to hypothetical situations. In this work, we focus on how people select situations that optimize specific goals for other people's emotions. We ask: to what extent can a predictive model of people's emotion be straightforwardly extended to do *planning*, instead?

Planning to evoke specific emotions necessarily involves rational choice over a model that can predict others' emotional reactions to potential interventions. There have been various proposals of how to model human-like predictions of others' emotions (Marsella et al., 2010; Ong et al., 2019), including rule-based emotion schema (Izard, 2007; Ortony et al., 1990), multi-agent computer simulations (Si et al., 2010; Alfonso et al., 2015; Yongsatianchot & Marsella, 2016), state-space transition dynamics (Thornton & Tamir, 2017), large language models (Rashkin et al., 2018; Sap et al., 2019), and intuitive theory-based probabilistic reasoning (Ong et al., 2015; Saxe & Houlihan, 2017; Wu et al., 2018; Houlihan et al., 2022).

The present research builds on recent modeling work that frames emotion prediction as causal reasoning over a Bayesian Theory of Mind (Houlihan et al., 2023; Ong et al., 2021). While each modeling approach has advantages, Bayesian Theory of Mind models present a distinct advantage in interpretability, granular reasoning, and generalization (Lake et al., 2017; Zhi-Xuan et al., 2022; Shu et al., 2021). Furthermore, by explicitly instantiating causal relations between situational features and mental states such as beliefs, desires, and emotions, they offer a finer degree of control over possible interventions that these models afford — a point we revisit in the discussion. We use the model of Houlihan et al. (2023) to simulate observers' predictions of other people's reactions to hypothetical situations, and apply it to the task of generating interventions to make an agent feel a target emotion. This is a hallmark of any causal model: the ability to generate interventions, simulate the effects, and compute counterfactuals (Pearl, 2009).

Using the model for planning provides a strong test for its generalization capabilities, by extending it to a task it was not initially designed or trained to do. Successful generalization would reinforce the validity of the model of human emotion understanding in general, and also suggest that causally structured generative models could support, and indeed might be critical for, a model of planning to change emotions.

Our approach of using strong generalization across tasks to underscore the value of causal models closely parallels previous arguments for causal models in intuitive physics (Battaglia et al., 2013), and is directly motivated by recent work in Theory of Mind (Ho, Saxe, & Cushman, 2022). Action understanding and Theory of Mind have often been

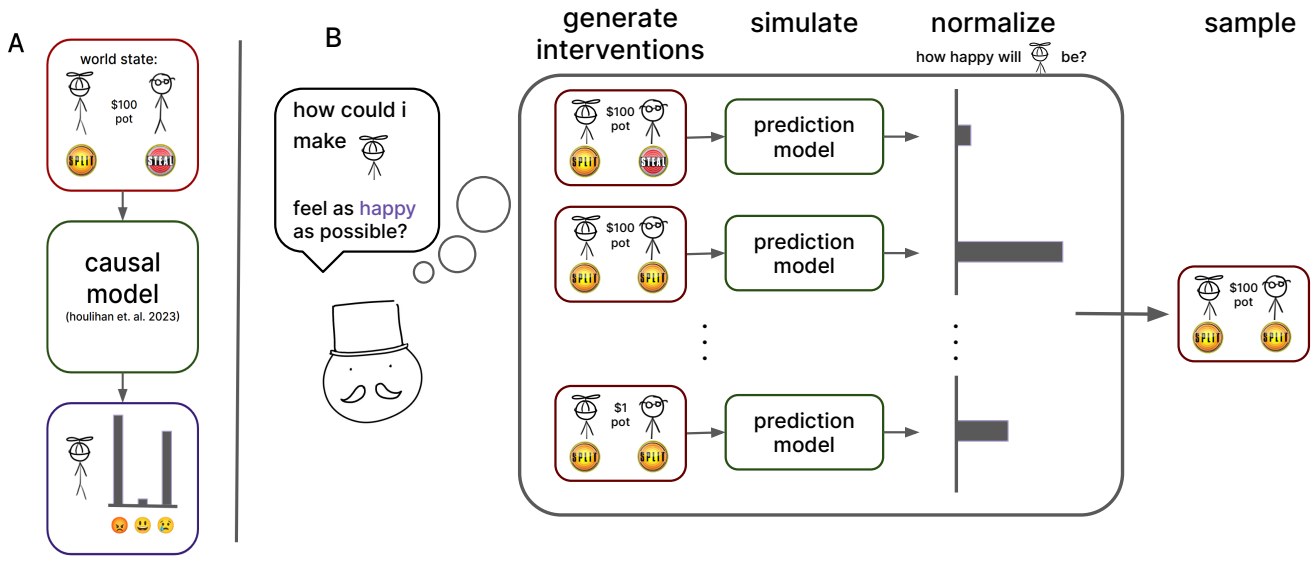

Figure 1: (A). The emotion prediction model of Houlihan et al. (2023). (B). Our proposed planning model. We generate interventions by simulating the effects of each intervention on the target emotion, normalizing all of the predicted intensities through a softmax function to obtain a probability distribution over interventions.

successfully modeled as *inverse planning*, where inferences about goals are obtained by inverting a causal model of how mental contents give rise to action (Baker et al., 2009; Jara-Ettinger et al., 2016; Gerstenberg & Tenenbaum, 2017). Bayesian inverse planning models have been used to account for human inferences about beliefs and desires (Baker et al., 2017), preferences (Jern et al., 2017), prosocial intent (Ullman et al., 2009), and moral judgments (Gerstenberg et al., 2018; Kleiman-Weiner et al., 2015). However, these models have been primarily used to fit human predictions about other people. Recent work has argued that our theories of Theory of Mind should be informed by functions beyond just prediction (Ho, Saxe, & Cushman, 2022), and our paper follows in a recent line of work using inverse planning models as models of social decision making, in domains such as pedagogy (Ho et al., 2017; Gweon, 2021) and impression maintenance (Yoon et al., 2020). Given that we are planning over an inverse planning-based model of emotions, our work can be construed as an application of *inverse inverse planning* (Chandra et al., 2023), which has been used to explain key features of human expression and depiction.

Using a causal generative model enables mental simulation of a range of hypotheticals and counterfactuals. In this work we model how people modify and craft scenarios to elicit specific emotional experiences in others, by asking them to plan over the situations that others encounter and the actions that others make. The underlying idea is that abstract reasoning over a generative Theory of Mind forms the computational basis for more constrained social cognition (such as planning how to modify an environment to change the emotions of a specific person, predicting how people will react to events, or inferring people's mental contents from their behavior).

Methods

Experimental setting

As mentioned, we are interested in interventions and decisions taken with the goal of inducing an emotion in an agent. To that end, we use the GoldenBalls game show as a test bed for all of our experiments and models. In GoldenBalls, a group of contestants play a sequence of games involving strategic deception and honesty. In the final segment, the only two players who have managed to avoid being eliminated play a variation of a one-shot prisoner's dilemma. Each player makes a choice to cooperate ("split") or defect ("steal") in private, and then the two players simultaneously reveal their choices. If both players choose "split" they each win half of the jackpot. If one player steals and the other splits, the player who chose "steal" wins the entire pot leaving the opposing player with nothing. If both steal, both players leave with nothing. All of our experimental paradigms are based on this implementation of the prisoner's dilemma.

The Split or Steal game implemented by GoldenBalls offers several experimental advantages. The game state can be described by a small number of variables, making modeling tractable, while the emotional variability present in the game is rich and naturalistic. The televised and highly-public nature of the game makes factors like reputation and humiliation much more salient than if the game were completely anonymous (although we collect emotion predictions for an anonymous variant in Experiment 2). Finally, GoldenBalls has been used in previous studies of emotion prediction (Houlihan et al., 2022, 2023), and our model directly builds on the work of

Houlihan et al. (2023).

To study planning for emotions, we focus on the interventions that the GoldenBalls game affords.The final round of a GoldenBalls episode can be represented as the actions of the two players, along with the amount of money at stake. Therefore, intervening on the GoldenBalls "world" consists of modifying the values of at most three variables. This gives us tractability in modeling while preserving some of the richness present in real-world social interactions.

Note that our interventions allow for directly manipulating a player's actions, which is not typically a naturalistic affordance. However, as previously discussed, people's capacity for abstract, hypothetical reasoning over a Theory of Mind is likely what enables more grounded action planning that generalizes well across a wide variety of social cognitive tasks.

Computational model

Houlihan et al. (2023) defined a model that takes in a specification of the game state (the actions of the players and the amount of money at stake) as input, and outputs a distribution of predicted emotion intensities conditional on that game state. Briefly, their model consisted of three sequential components designed to emulate how observers predict players' emotional reactions. Component (1) uses inverse planning to infer what mental contents were likely to have motivated a player's action. This yields a joint distribution over a player's beliefs about their opponent, inferred reputational consequences for acting prosocially or competitively, preferences for selfish and social outcomes, and desires to manage other people's inferences. Component (2) computes a joint distribution over *appraisal variables* that encode evaluative features of the situation, combining the outcome of the game with the beliefs and preferences of the player inferred via inverse planning. Each appraisal variable represents an expected, achieved, or counterfactual utility for monetary features, first-order social features such as fairness and inequity aversion, or higher-order reputational features such as the player's inference of how much people observing the game will believe that the player's action was motivated by selfish or prosocial goals. Component (3) transforms the computed distribution over appraisal variables into predicted intensities of 20 emotions.

More formally, their prediction model infers a joint distribution $p(e \mid a_1, a_2, j)$, where $e \in [0,1]^{20}$ is the intensity of 20 target emotions that the target agent (player 1) is expected to feel (see Figure 3 for the full list of emotion labels), $a_1 \in \{\text{split, steal}\}$ and $a_2 \in \{\text{split, steal}\}$ are the actions of the target player and the opposing player, respectively, and j is the size of the jackpot. Since we are only interested in a target emotion e_t, we obtain the distribution $p(e_t|a_1, a_2, j)$ by marginalizing over all of the other emotions. We note that for more sophisticated planning that involves maximizing or minimizing several emotions jointly, another approach should be considered that preserves the covariance between predicted emotions.

Importantly, their model was designed and trained exclusively for prediction, not planning interventions. The novel contribution of the current work is to use the prediction model (Figure 1A) as a way to simulate the effects of interventions, in service of selecting the intervention that best produces the desired emotional state (Figure 1B). The expected effect of intervention $I = (a_1, a_2, j)$ on e_t is given by $\mathbf{E}[e_t \mid a_1, a_2, j]$. This expectation is calculated by assuming the game state and player actions are defined according to I, running the predictive model forward to generate predicted emotion intensities, and then averaging over the predictions.

We use the softmax choice function (Luce, 2005) to decide between interventions. That is, for each intervention I, our model selects that intervention with probability $p(I) \propto \exp(\beta \cdot \mathbf{E}[e_t|I])\, p_0(I)$, where β is the inverse temperature parameter that controls the "rationality" of the decision maker. We choose $p_0(I)$ to be the uniform distribution over interventions.

We emphasize that while our model would work with any black box emotion predictor, our overall hypothesis is that emotion intervention is built on an intuitive theory of others' emotions that is abstract, generative, and causally structured, just as human Theory of Mind is.

Baselines While our goal is to understand whether a causal model of emotions can be applied to planning interventions, it's possible that not all components of the model are necessary for planning. As such, we consider two alternative baselines, corresponding to ablations of the full model.

The social lesioned model is a model that predicts interventions solely based on two monetary features: achieved utility (how much money was won or lost by the target player), and prediction error (how much money was won or lost by the target player relative to their expected payout), thus ignoring social values (e.g. inequity aversion) and reputational factors. We implement this model by applying the same planning procedure to the SocialLesion model of (Houlihan et al., 2023).

The second is a purely statistical model (the Unconditional model) that samples interventions from a prior $p(I) = p(a_1) \cdot p(a_2) \cdot p(J)$, without conditioning on the particular emotion. We set each of these prior probabilities to empirical frequencies calculated from Experiment 2. The purpose of this baseline is to test for variability in interventions between emotions: if people choose the same interventions irrespective of which emotion they were trying to elicit, then we would expect this model to account well for human judgments.

Experiment 1

Previous work has shown that participants have strong intuitions about the emotions that specific outcomes in Goldenballs will elicit (Houlihan et al., 2023). Our first preregistered experiment instead asks them to leverage these intuitions to plan interventions to elicit a target emotion, to test whether the choices they make align with predictions made by our model[1]. To do so, we first focus on the simplest possible in-

[1]Preregistered at: https://aspredicted.org/j2mc4.pdf

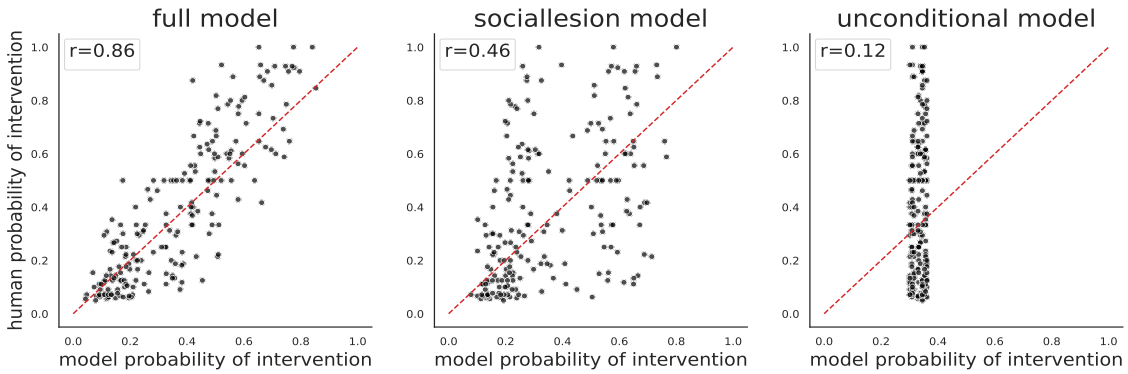

Figure 2: Model predictions versus human choices in Experiment 1 for (A) our proposed model, (B) a model that makes interventions based solely on the monetary features, and (C) a model that chooses interventions based on empirical frequencies collected from humans. For each target emotion to be elicited and original game state, we plot the probability that the model assigns to any given intervention on that original game state, against how likely humans are to select the same intervention. Our model significantly outperforms all other models in accounting for player interventions.

tervention: changing only one player's actions.

Procedure Participants were introduced to the structure of the GoldenBalls game. They were told that they would first see the original result of a round of GoldenBalls, and then indicate what the players should have done differently, in order to make a target player feel more of a particular emotion. They were given three options: to change what action the target player chose (flipping from steal to split or vice versa), what action the opposing player chose, or to leave both players' choices the same. Crucially, they were not allowed to manipulate both player's choices simultaneously. The exact value of the jackpot, and the original decisions of the players were randomized for each trial. Following previous work, we defined the jackpot sizes to be $\{\$77, \$124, \$61{,}430, \$138{,}238\}$.

On each trial, participants were shown the target player and the opposing player, instructed on which emotion they were tasked to elicit, and were shown the original outcome of the game. They then proceeded to choose an intervention out of the three offered. We used a subset of 8 face stimuli for the players from Houlihan et al. (2023), and randomized the faces across trials so that no participant saw the same face twice.

We probed for interventions on 20 emotion labels (see Figure 3). Each target emotion was associated with four possible original game states (whether each player "split" or "stole") that participants were asked to intervene on, resulting in a total of 80 stimuli.

Participants We recruited participants (n=300) from Prolific. The task took approximately 6 minutes, for which participants were compensated \$1.50. We excluded the fastest 10% of participants from our analyses.

Results For each original game state $S = (a_1, a_2, j)$, there are three possible choices: S itself (corresponding to no intervention), the state with the target player's action changed: $(\neg a_1, a_2, j)$, and the state with the opposing player's action changed: $(a_1, \neg a_2, j)$. The planning model uses the softmax function so that the probabilities of choosing each of these three interventions sum to 1. We select the value of β that best fits the data through a random grid search procedure. All subsequent analyses and experiments are run with β frozen.

Results are shown in Figure 2. Empirically, we find that the model-predicted probability of an intervention aligns closely with the empirical human probability of choosing that intervention ($r = 0.86$). Furthermore, our model drastically outperforms the two baselines on this same task ($r = 0.46, 0.12$), respectively for the SocialLesion and Unconditional models.

We fit multinomial mixed effects regressions with random intercepts by participant, regressing the choice of intervention against the probability that the model chooses that same intervention. We find a significant effect for our model ($\beta = 2.36,\ \text{SE} = 0.07,\ p < 10^{-16}$).

Experiment 2

Experiment 1 showed that participant's choices of what interventions to take were well predicted by our computational model. However, the space of interventions was heavily restricted, only allowing for one of the player's actions to be modified.

To that end, in our second preregistered experiment, we instead ask participants to *design* world states to elicit a target emotion[2]. Instead of acting to change a particular world state, participants are now asked to create entirely new states. This not only offers a conceptual replication and extension of Experiment 1, it also allows us to test the generalization capabilities of our model, by fixing the sole free parameter β and generalizing to newly collected data.

Experiment 1 showed that there is a substantial amount of between-emotion variability in the interventions that people choose. However, the state space of the game (players' actions and jackpot size) is too low dimensional to separate out

[2]Preregistered at https://aspredicted.org/z3sa8.pdf

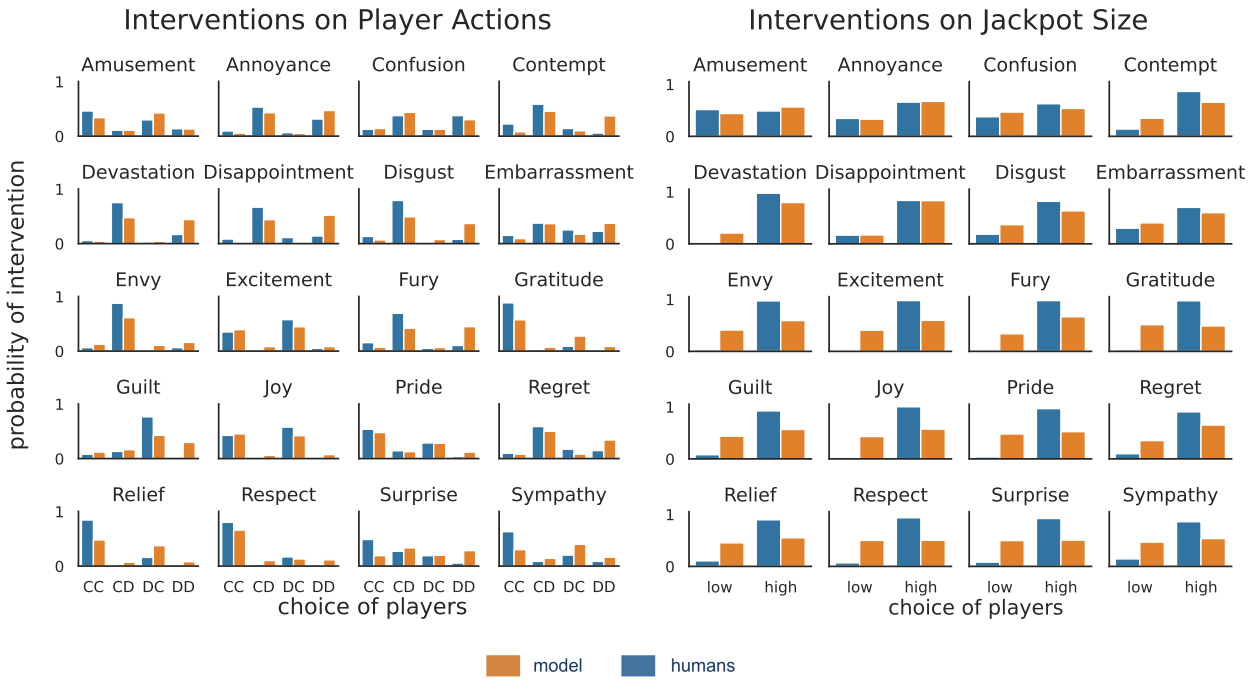

Figure 3: Distributions for interventions on player actions (left) and jackpot size (right) for Experiment 2. We use C to represent cooperation and D for defection in the prisoner's dilemma. These distributions were obtained by marginalizing over all other variables, and the height of each bar indicates the marginal probability that the model or humans assign to that specific choice. Our model accounts for most of the patterns seen in interventions on player choice, but notably, remains ambivalent about jackpot size in cases where human participants almost unanimously choose to make the jackpot size large.

all of the 20 emotions that we probe for. To further distinguish between intervening on emotions such as *Embarrassment* as opposed to *Regret*, we introduce a new variable that allows participants to vary the reputational structure of the game. Participants are introduced to two variants of GoldenBalls — the public game (i.e. the variant shown in exp 1), in which players engage in the prisoner's dilemma on live television and a studio audience, and the private game, in which the game is played over the radio with players calling into the station and communicating exclusively over a text interface. Thus, in the private game, the players' identities are completely obscured from both the audience and each other.

Procedure Participants were first introduced to the structure of the GoldenBalls game. They were told that the task was to design specific scenarios to make a player feel a target emotion, by selecting (1) the action of the target player (whether the player split or stole) (2) the action of the opposing player, (3) the amount of money at stake (whether the jackpot should be high or low), and (4) the reputational structure of the game (whether the game should be public or private). The exact values of the low / high jackpot size were randomized between trials. We split the jackpot sizes used in Experiment 1 into two groups: the low jackpot sizes were $\{\$77, \$124\}$ and the high jackpot sizes were $\{\$61,430, \$138,238\}$.

On each trial, participants were shown the target player and opposing player, instructed as to which emotion to elicit, and given the possible interventions to choose from. We probe the same 20 emotions as in Experiment 1, and the identities of each player were randomized between trials in the same fashion.

Participants We recruited participants (n=250) from the Prolific research platform. The task took approximately 6 minutes, for which participants were compensated $1.50. We excluded the fastest 10% of participants from our analyses.

Results All predictions were made with the value of β learned in Experiment 1. To generate model predictions for interventions, we plan over all 8 game states (a_1, a_2, j). Note that the emotion prediction model was only trained to predict human judgments for the original (i.e. public) version of GoldenBalls. At present, we do not generate model predictions for the reputational structure manipulation, but rather restrict modeling to the jackpot size and players' actions.[3]

Results are shown in Figures 3 and 4. As with Experiment 1, our model closely predicts the interventions that humans make on the player choices. To quantify this fit,

[3]However, these reputational interventions may serve as useful data for future elaborations of our model.

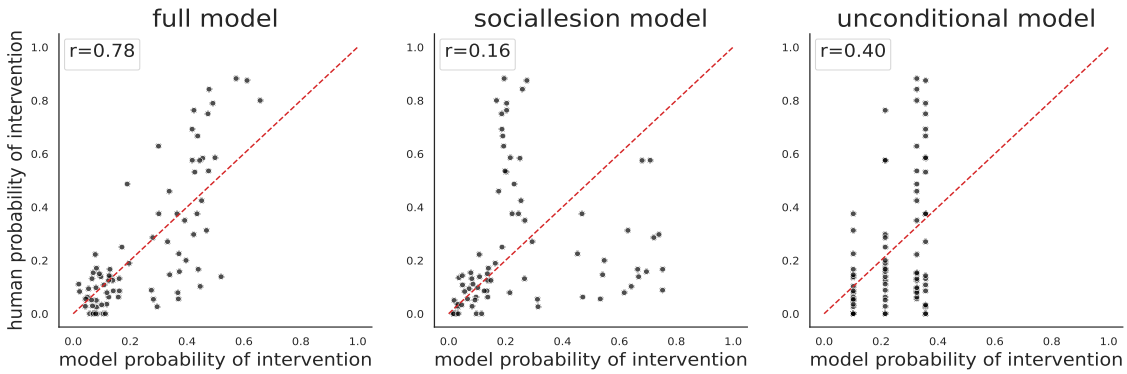

Figure 4: Model and human predictions for the player action interventions of Experiment 2. The plots show the model probability versus human probability of selecting an intervention for (A) our proposed model, (B) a model that makes interventions based on the monetary gain or loss for the target player, and (C) a model that chooses interventions according to a prior distribution, not conditioned on the particular emotion. Our model significantly outperforms all other models in accounting for player interventions.

we compute the KL-Divergence $D_{\mathrm{KL}}(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$ between the distribution over interventions predicted by the model and by humans for each emotion. We find that our model shows the closest average alignment with the empirical distribution over interventions for humans, with $\mathrm{KL} = 0.049$, dramatically outperforming the SocialLesion model and Unconditional model ($\mathrm{KL} = 0.12, 0.29$, respectively).

One key area of divergence between our model and people is that people almost universally choose to set the jackpot size to be high, but our model remains non-committal about jackpot size for certain emotions. This is because the emotion prediction model infers that these particular emotions depend less on the size of the jackpot and more on other features of the game, and so varying the size of the jackpot only slightly influences the predicted emotion intensity.

Discussion and Future Directions

Our results suggest that when people are asked to design interventions that affect others' emotions, their choices are broadly consistent with model-based planning over a causally-structured generative Theory of Mind. At the same time, there still remains a substantial explanatory gap between the model and human data. Is this gap in performance due to the emotion prediction model, or in the decision rule built on top of the predictions? Future work should investigate the extent to which deviations from human planning can be mitigated by simply training a better prediction model—or whether jointly training the model to do both prediction and planning simultaneously is needed.

Because our goal was to test how a model exclusively trained on prediction might generalize to planning, the model has no explicit representation of goals, or is goal-directed in any way. However, recent work has emphasized the goal-directed nature of our representations (Ho, Abel, et al., 2022), suggesting that our models adapt to the task at hand. As such, we might expect that our internal predictive model is not only able to predict people's emotions and behavior, but also integrate additional information or discard irrelevant information when necessary, depending on the goal.

Our results support the conclusion that people in our task are planning interventions by leveraging a causal model of emotions with structured representations (such as beliefs and desires), but they do not rule out other views regarding how people mentally represent and use social information. While the emotion prediction model employed in this work uses structured latent representations, we treat the model as a "black box", and could theoretically obtain the same results from a less structured model that does not explicitly represent mental contents such as beliefs, desires, and appraisals. Further evidence could come from more aggressive tests of generalization, asking people to intervene on situations that the prediction model was not explicitly trained on. Another avenue is to test interventions on intermediate states of the model, such as interventions on agents' beliefs or desires. Doing so would distinguish models with explicit representations of beliefs and desires, without which these interventions would not be possible. In general, a promising future direction is to work towards further evidence for *how* people predict or plan, by establishing or breaking links in the causal chain of the model.

Finally, in this work we primarily focused on targeted interventions in a limited domain, mostly consisting of changing the values of several binary variables. However, much of human expression and storytelling involves constructing worlds that are considerably less constrained. In storytelling, the only limiting factor in the space of interventions is the imagination of the author, but for it to be a good story, it should still adhere to people's abstract and causal intuitions connecting situations with their expected emotional experiences. A promising direction of future work is to extend our model to model aspects of human storytelling, depiction, and acting (Chandra et al., 2023).

Acknowledgements

The authors would like to thank the anonymous reviewers for helpful feedback and comments. TC is supported by an NDSEG fellowship, DH is supported by a fellowship from the William H. Neukom Institute for Computational Science, KC is supported by the Hertz Foundation and the NSF GRFP under grant #1745302, and RS is supported by the Paul E. and Lilah Newton Brain Science Award, and a Center for Brains, Minds and Machines grant funded by NSF STC award CCF-1231216.

References

Alfonso, B., Pynadath, D. V., Lhommet, M., & Marsella, S. (2015, September). Emotional perception for updating agents' beliefs. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 201–207). Xi'an, China: IEEE. doi: 10.1109/ACII.2015.7344572

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017, March). 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, December). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. doi: 10.1016/j.cognition.2009.07.005

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Chandra, K., Li, T.-M., Tenenbaum, J. B., & Ragan-Kelley, J. (2023). Storytelling as inverse inverse planning. *Topics in Cognitive Science*.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive Theories. In M. Waldmannn (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? from expectations to responsibility judgments. *Cognition*, *177*, 122–141.

Gweon, H. (2021, October). Inferential social learning: cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910. doi: 10.1016/j.tics.2021.07.008

Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2022). People construct simplified mental representations to plan. *Nature*, *606*(7912), 129–136.

Ho, M. K., MacGlashan, J., Littman, M. L., & Cushman, F. (2017, October). Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition*, *167*, 91–106. doi: 10.1016/j.cognition.2017.03.006

Ho, M. K., Saxe, R., & Cushman, F. (2022, September). Planning with Theory of Mind. *Trends in Cognitive Sciences*. doi: 10.1016/j.tics.2022.08.003

Houlihan, S. D., Kleiman-Weiner, M., Hewitt, L. B., Tenenbaum, J. B., & Saxe, R. (2023, July). Emotion prediction as computation over a generative theory of mind. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *381*(2251), 20220047. doi: 10.1098/rsta.2022.0047

Houlihan, S. D., Ong, D., Cusimano, M., & Saxe, R. (2022). Reasoning about the antecedents of emotions: Bayesian causal inference over an intuitive theory of mind. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, pp. 854–861).

Izard, C. E. (2007, September). Basic Emotions, Natural Kinds, Emotion Schemas, and a New Paradigm. *Perspectives on Psychological Science*, *2*(3), 260–280. doi: 10.1111/j.1745-6916.2007.00044.x

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016, August). The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604. doi: 10.1016/j.tics.2016.05.011

Jern, A., Lucas, C. G., & Kemp, C. (2017, November). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64. doi: 10.1016/j.cognition.2017.06.017

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of Intention and Permissibility in Moral Decision Making. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 37, pp. 1123–1128).

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

Luce, R. D. (2005). *Individual choice behavior: A theoretical analysis*. Courier Corporation.

Marsella, S., Gratch, J., & Petta, P. (2010). Computational models of emotion. In K. R. Scherer, T. Bänziger, & E. Roesch (Eds.), *A Blueprint for Affective Computing-A sourcebook and manual* (pp. 21–46). Oxford University Press.

Nelissen, R., & Zeelenberg, M. (2009). When guilt evokes self-punishment: evidence for the existence of a dobby effect. *Emotion*, *9*(1), 118.

Ong, D. C., Soh, H., Zaki, J., & Goodman, N. D. (2021, April). Applying Probabilistic Programming to Affective Computing. *IEEE Transactions on Affective Computing*, *12*(2), 306–317. doi: 10.1109/TAFFC.2019.2905211

Ong, D. C., Zaki, J., & Goodman, N. D. (2015, October). Affective cognition: Exploring lay theories of emotion. *Cognition*, *143*, 141–162. doi: 10.1016/j.cognition.2015.06.010

Ong, D. C., Zaki, J., & Goodman, N. D. (2019, April). Computational models of emotion inference in Theory of Mind: A review and roadmap. *Topics in Cognitive Science*, *11*(2), 338–357. doi: 10.1111/tops.12371

Ortony, A., Clore, G. L., & Collins, A. (1990). *The Cognitive Structure of Emotions*. Cambridge University Press.

Pearl, J. (2009). *Causality*. Cambridge university press.

Rashkin, H., Sap, M., Allaway, E., Smith, N. A., & Choi, Y. (2018). Event2Mind: Commonsense Inference on Events, Intents, and Reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 463–473). Melbourne, Australia: Association for Computational Linguistics. doi: 10.18653/v1/P18-1043

Sap, M., Rashkin, H., Chen, D., Le Bras, R., & Choi, Y. (2019, November). Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 4463–4473). Hong Kong, China: Association for Computational Linguistics. doi: 10.18653/v1/D19-1454

Saxe, R., & Houlihan, S. D. (2017, October). Formalizing emotion concepts within a Bayesian model of theory of mind. *Current Opinion in Psychology*, *17*, 15–21. doi: 10.1016/j.copsyc.2017.04.019

Shu, T., Bhandwaldar, A., Gan, C., Smith, K., Liu, S., Gutfreund, D., … Ullman, T. (2021). AGENT: A benchmark for core psychological reasoning. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning* (Vol. 139, pp. 9614–9625). PMLR.

Si, M., Marsella, S. C., & Pynadath, D. V. (2010, January). Modeling appraisal in theory of mind reasoning. *Autonomous Agents and Multi-Agent Systems*, *20*(1), 14–31. doi: 10.1007/s10458-009-9093-x

Thoits, P. A. (1996, June). Managing the Emotions of Others. *Symbolic Interaction*, *19*(2), 85–109. doi: 10.1525/si.1996.19.2.85

Thornton, M. A., & Tamir, D. I. (2017, June). Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences*, *114*(23), 5982–5987. doi: 10.1073/pnas.1616056114

Tran, A., Greenaway, K. H., Kostopoulos, J., O'Brien, S. T., & Kalokerinos, E. K. (2023, December). Mapping Interpersonal Emotion Regulation in Everyday Life. *Affective Science*, *4*(4), 672–683. doi: 10.1007/s42761-023-00223-z

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. (2009). Help or Hinder: Bayesian Models of Social Goal Inference. In *Advances in Neural Information Processing Systems* (Vol. 22). Curran Associates, Inc.

Wu, Y., Baker, C. L., Tenenbaum, J. B., & Schulz, L. E. (2018). Rational Inference of Beliefs and Desires From Emotional Expressions. *Cognitive Science*, *42*(3), 850–884. doi: 10.1111/cogs.12548

Wu, Y., Schulz, L. E., Frank, M. C., & Gweon, H. (2021). Emotion as information in early social learning. *Current Directions in Psychological Science*, *30*(6), 468–475.

Yongsatianchot, N., & Marsella, S. (2016). Integrating Model-Based Prediction and Facial Expressions in the Perception of Emotion. In B. Steunebrink, P. Wang, & B. Goertzel (Eds.), *Artificial General Intelligence* (pp. 234–243). Cham: Springer International Publishing.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020, November). Polite Speech Emerges From Competing Social Goals. *Open Mind*, *4*, 71–87. doi: 10.1162/opmi_a_00035

Zhi-Xuan, T., Gothoskar, N., Pollok, F., Gutfreund, D., Tenenbaum, J. B., & Mansinghka, V. K. (2022, August). *Solving the Baby Intuitions Benchmark with a Hierarchically Bayesian Theory of Mind.* arXiv. doi: 10.48550/ARXIV.2208.02914