Action understanding with presentational goals

Victor Btesh[1*], **Sarah Wu**[2], **Tobias Gerstenberg**[2]
[1]University College London
[2]Stanford University
[*]vbtesh9@gmail.com

Abstract

Humans balance many goals while navigating social interactions. We are motivated not only by what we want for ourselves and for others, but also by what we want others to think about us. Inferences about agents' personal and social goals have been formally captured using Bayesian inverse planning models. However, less is known about the mechanisms underlying our ability to infer *presentational goals*: an agent's desires over how another agent sees them. We introduce a novel paradigm that allows us to test joint inferences about social and presentational goals. We propose an extension to inverse planning where agents derive additional utility from shifting others' beliefs about their social goals. Because the model places no constraint on the valence of presentational targets, it naturally captures cases where agents wish to appear prosocial or adversarial. Across a variety of scenarios, participants make systematic joint inferences about social goals, presentational goals, and presentational targets. Our computational model captures participants' inferences better than feature-based alternatives.

Keywords: theory of mind; action understanding; social cognition; Bayesian inference; presentational goals

Introduction

Imagine yourself in a park. You glance at a play area and see two children, Alice and Bob. Alice teases Bob and takes his toy despite his protestations. You drift off in a book and lift your head a moment later as Charlie, the parent of Alice and Bob, arrives and waves at you. As you turn your gaze back towards Alice and Bob, you see her helping Bob with his toy, carefully caring for him. You gently smile at this sight, noticing Alice's efforts to please Charlie, without thinking much of it. Yet, the inference you just drew is anything but trivial. It rests on rich assumptions about what motivates us in social relationships, such as presentation of self and reputation.

Humans are able to draw rich inferences from sparse data when interpreting the behaviour of others. Previous research suggests that, to achieve this, people invert a generative model of what a rational agent would do to determine others' goals and preferences (Baker et al., 2009, 2017). This perspective offers a powerful explanation for how we draw inferences from little evidence by proposing that we leverage a robust yet naive generative model of what motivates people: they act to maximise their utility (Jara-Ettinger et al., 2016).

Early work focused on agents with solely self-interested goals, such as reaching a desired location or object. More recent research has incorporated social goals that involve adopting another agent's utility into one's own (Baker et al., 2008; Powell, 2022; Ullman et al., 2009). Inverse planning models account well for people's intuitive inferences about these kinds of social goals (Netanyahu et al., 2021; Shu et al., 2020; Tejwani et al., 2022). Not only are people apt at making inferences about whether an agent wants to harm or help another, but they can also use them as input to richer moral judgments such as responsibility and blame (Wu et al., 2023).

Yet, navigating social interactions and maintaining interpersonal relationships can often involve even higher order goals over the mental states of others. Impression management involves holding preferences over others' beliefs of ourselves, e.g. we may want others to think we are prosocial, inequity averse, or competent (Goffman, 1959; Kleiman-Weiner et al., 2017; Zhao et al., 2021). Children from a young age hold preferences about what adults think of their competence and attempt targeted demonstrations of ability to move the beliefs of adults in desirable states (Gweon, 2021) and are sensitive to socially-mindful actions such as leaving a choice for others (Zhao et al., 2021). Crucially, people happily trade these objectives off with material outcomes by signalling an intention to collaborate by forgoing immediate reward (Battigalli & Dufwenberg, 2022; Roberts, 2020). Computational accounts of such higher order goals have had to take seriously the idea that targeted changes in the belief states of others can constitute meaningful sources of reward for people, thus naturally drawing conceptual links to rational theories of communication (Chandra et al., 2024; Degen, 2023; Yoon et al., 2020).

The present work focuses on higher order goals defined over what other people believe about our social goals. Recent research has proposed that presentational goals, that is, attempting to control the impression we give to others, may be formulated as utility derived from maintaining the beliefs others have of how much we weigh different rewards when taking an action (Houlihan et al., 2023; Kleiman-Weiner et al., 2017). This view proved useful in showing that people are sensitive to how ready others are to change their mind when considering presentational interventions. This leads them to act more selfishly when they feel others are convinced of their good intentions (Btesh et al., 2025).

Here, we study inferences about agents' social and presentational goals in a novel experimental paradigm that allows us to pit them directly against each other, and also against personal goals. These three types of goals may not always be

Figure 1: **An example story** (story C in Figure 2) as it was presented to participants with the questions that participants were asked to answer. We expected participants to interpret Red as trying to convince Purple that they like Green. **1.** Red and Green both prefer apples. Purple sees them choose and now knows what they each prefer. Red also knows what Green prefers. **2.** Red and Green wait in line. Red goes first and takes the apple. Purple is absent. **3.** Red and Green wait in line. Red goes first and takes the orange. Purple sees them choose.

aligned, which can produce counter-intuitive yet informative behaviour. Nevertheless, we find that people are able to make systematic joint inferences from just a few observations, and we develop a computational model that extends the naive utility calculus and accurately captures patterns of participants' judgments.

Experimental paradigm

We introduce a new theory of mind task to systematically assess people's intuitive theories about how people trade off personal, social, and presentational goals. This task allows for these goals to come apart—for example, when Alice may not like Bob but wants Charlie to think she does.

Participants saw 16 stories involving three characters: Red (the actor), Green (affected by Red's action), and Purple (a third party observer). Each story comprised three scenes. The first scene showed Red and Green independently choosing between an apple and an orange, with Purple being present, too. This scene created common knowledge about which fruit each character prefers. The second scene showed Red and Green waiting in line to pick a fruit. Red always picked first and chose between either an apple and an orange, or two apples and an orange, leaving Green with the remaining fruit(s). Green's choice afterwards was not shown. This scene thus provided information about Red's social goal towards Green. The third scene shares the structure of the second, but now Purple is back, thus providing an opportunity to observe whether Red's behaviour changes in the presence of Purple. For instance, in Figure 1, both Red and Green prefer apples over oranges (scene 1). Red takes the apple when Purple is absent (scene 2), but takes the orange when Purple is watching (scene 3), signalling adoption of Green's utility. After the three scenes, participants answered questions about Red's social and presentational goals (see Figure 1).

Computational model

We assumed that participants would reason about Red's behaviour by inverting a generative model for what motivates Red. Red weighs three goals when choosing a fruit: a personal goal (what they prefer), a social goal (how their choice affects Green), and a presentational goal (how this choice may shift Purple's assessment of Red's social goal). Following Btesh et al. (2025), the action that Red takes changes Purple's belief about Red, and a targeted change in Purple's belief is itself a source of utility.

Personal goal Given a fruit $f \in \{\text{apple}, \text{orange}\}$, Red and Green derive reward $V_i(f, \alpha_i) = \alpha_i$ if $f = \text{apple}$ and $1 - \alpha_i$ otherwise, where $\alpha_i \in [0, 1]$ weights how much agent i prefers apples. Red thus places self-interested utility on each fruit f:

$$U_{\text{personal}}(f) = V_{\text{red}}(f, \alpha_{\text{red}}). \tag{1}$$

Social goal A socially motivated Red additionally weighs Green's reward. Formalising this as adopted utility (Powell, 2022; Ullman et al., 2009; Wu et al., 2023), Red additionally places social utility:

$$U_{\text{social}}(f, \beta) = -\beta \cdot V_{\text{green}}(f, \alpha_{\text{green}}) \cdot L_{\text{green}}(k, n_f) \tag{2}$$

on each fruit f, where $\beta \in [-1, 1]$ controls how much Red weighs Green's utility and α_{green} encodes Green's preference. The negative sign in front of β represents the zero-sum nature of the setting: Red's choice f is a loss for Green if Green also prefers f. $L_{\text{green}}(k, n_f) \in \{0, 1\}$ captures how many of Green's desired fruits Red's choice removes from the pool. In practice $L_{\text{green}}(k, n_f) = \min(1, \max(0, k - n_f + 1))$, where n_f is the count of the fruit Red picked (before Red's action) and $k \in \{1, 2\}$ is the number of fruits Green would prefer to eat. This lets Green incur no loss should two apples be present if $k = 1$, or a loss of one otherwise. An agent considering both goals acts according to:

$$U_{\text{simple}}(f, \beta) = U_{\text{personal}}(f) + U_{\text{social}}(f, \beta) \tag{3}$$

Presentational goal A Red who cares about how Purple perceives them, additionally derives reward from shifting Purple's beliefs. Let $p_{\text{purple}}(\beta | f)$ denote the posterior Purple would hold over Red's prosociality given that Red picked fruit

f and let $\rho(\beta_{\text{goal}}) = \mathcal{N}(\beta_{\text{goal}}, \sigma^2_{\text{goal}})$ be the belief Red wants Purple to hold. This produces presentational utility

$$U_{\text{presentational}}(f, \delta, \beta_{\text{goal}}) = -\delta \cdot D_{KL}[p_{\text{purple}}(\beta|f)||\rho(\beta_{\text{goal}})] \quad (4)$$

where $\delta \in [0, 1]$ controls how much Red cares that Purple's belief matches their target and the KL term measures the gap between both. Red therefore benefits from actions which move Purple's beliefs towards $\rho(\beta_{\text{goal}})$. Three features of this formulation are worth highlighting. First, δ and β_{goal} are conceptually distinct: δ controls whether Red has a goal over Purple's beliefs at all while β_{goal} is concerned about what this goal is. Second, the model is valence agnostic: β_{goal} ranges over $[-1, 1]$, the same as β, allowing Red to want to appear prosocial, neutral, or adversarial. Third, when Purple is absent, no belief update is possible and thus $p_{\text{purple}}(\beta|f) = p_{\text{purple}}(\beta)$, making the KL term constant and removing any presentational contribution to Red's actions (Btesh et al., 2025). Intuitively, presentational utility towards Purple should only bear on action selection when Purple is watching, so all identifying information about δ comes from scene 3. Putting all three goals together, Red chooses fruit f with utility function:

$$\begin{aligned} U(f, \beta, \delta, \beta_{\text{goal}}) = U_{\text{personal}}(f) + U_{\text{social}}(f, \beta) \\ + U_{\text{presentational}}(f, \delta, \beta_{\text{goal}}) \end{aligned} \quad (5)$$

Beliefs

Three perspectives are at play in interpreting each story. First, the observer (or participant), who is uncertain about Red's preferences and about Red's beliefs regarding Green and Purple. Second, Red, who is uncertain about Green. Third, Purple (as represented by Red), who is uncertain about Red and Green's preferences. Throughout, we model agent i's commitment to the preferences revealed in scene 1 as $p(\alpha_i|f_i) = \text{Beta}\big(\tilde{f}_i \cdot \tau, (1 - \tilde{f}_i) \cdot \tau\big)$, where $\tilde{f}_i \in \{\epsilon, 1 - \epsilon\}$ with $\epsilon = 0.05$ is a smoothed indicator of i's choice (0 for apple, 1 for orange) preventing evaluating Beta at 0 or 1 and τ is a precision parameter acting as Beta pseudo-counts.

Observer's beliefs (participants) The observer assumes Red acts according to Equation 5 and holds a uniform prior over $p(\beta, \delta, \beta_{\text{goal}})$. Beyond these three parameters, the observer is uncertain about how committed Red is to their scene 1 preference, modelled as $p(\alpha_{\text{red}}|f_{\text{red}})$ with precision τ_{obs}.

Red's beliefs about Green. As Green does not act, they are only represented in the adopted term of Red's utility. Red holds the analogous belief $p(\alpha_{\text{green}}|f_{\text{green}})$ with precision τ (distinct from τ_{obs}). Red is also uncertain about how many of their preferred fruits Green would want, $k \in \{1, 2\}$, parametrised by $\theta = p(k = 2)$. This matters only in scenes where there are two apples and one orange.

Red's model of Purple's beliefs Evaluating the presentational term requires that Red holds a model of how Purple interprets Red. Following Btesh et al. (2025), we assume Red takes Purple to model Red as acting with personal and social goals only, i.e. according to Equation 3. Purple's prior over β is uniform, and their beliefs about both agents' scene 1 preferences follow the construction above with precision τ. It matches the precision Red uses for their own beliefs about Green, since Red knows Purple observed scene 1. Red then applies Bayes' rule to compute the posterior distribution $p_{\text{purple}}(\beta|f) = \sum_{\alpha_{\text{red}}} p_{\text{purple}}(\alpha_{\text{red}}, \beta|f)$ that Purple *would have* if Red picked f after marginalising out Red's preferences. We additionally let Red hold presentational goals toward Green following the same structure, but marginalise them out in the reported posterior over Red's goals toward Purple.

Conditioned on the number of fruits, preferences revealed in scene 1, and Red's subsequent choices f_1 and f_2, the model returns a joint posterior distribution over β, δ, β_{goal}, marginalising over α_{red}, α_{green} and k and any presentational goals Red might hold toward Green.

Model variants

Our prediction is that participants answer each question by assuming that Red chooses fruits f_1 and f_2 according to our proposed generative model and invert it to infer a joint posterior over each parameter: $p(\beta, \delta, \beta_{\text{goal}}|f_1, f_2)$.

Marginals from the joint posterior The most natural way to generate a response is to sample from each parameter's marginal distribution. This constitutes our first model. This model has five parameters: the precision of the observer τ_{obs}, Red's precision τ, the probability that Green eats two rather than one apple θ, the precision of the presentational target $\rho(\beta_{\text{goal}})$ denoted σ_{goal}, and finally an inverse temperature parameter T^{-1} for a softmax representing how utility-rational participants assume Red to be when choosing their fruit.

Estimating δ by conditioning on $\beta_{\text{goal}} = 1$ Our second variant modifies how δ is estimated. Rather than reading it directly from the marginal posterior $p(\delta|f_1, f_2)$, this variant reads it from the conditional $p(\delta|f_1, f_2, \beta_{\text{goal}} = 1)$. This corresponds to interpreting "how much does Red care what Purple thinks of Red" as "How much does Red care that Purple thinks they like Green". This second model shares all parameters with the first one.

Heuristic for δ Our third variant mixes the conditional estimate of δ from the previous variant with a simple behaviour change rule: $\delta = 1$ if Red's choice differs when Purple watches, i.e. between scenes 2 and 3, and $\delta = 0$ otherwise. Two parameters govern the mixture: a weight α controlling the proportion of people who answer the question in this way and the noise in those participants' responses $\sigma_{\text{heuristic}}$, yielding a total of seven parameters for this last model.

Feature-based model To test whether participants were answering each question using the features of each story rather than doing inverse planning, we fit three linear mixed effects models predicting β, δ, and β_{goal} from story data. Predictors

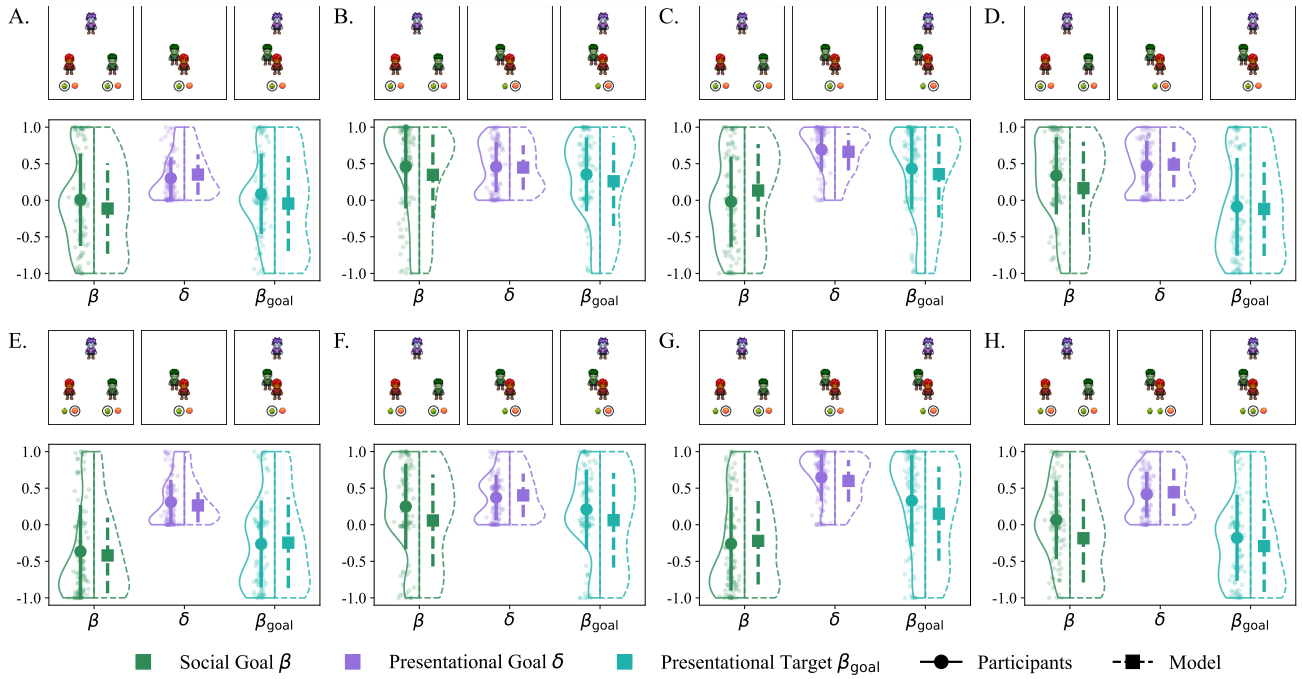

Figure 2: **Distributions of participants and model responses for selected stories**. The three images at the top of each plot depict each scene of the story being modelled. Violin plots represent the posterior marginal distributions for each judgment; points represent means and error bars are standard deviations. Full lines are participants and dotted lines are model predictions for the model that conditions on β_{goal} and uses the heuristic when Red changes their behaviour in the presence of Purple.

included binary encodings of whether preferences were competitive or complementary, the number of apples (i.e. 2 or 3), Red's first choice f_1, Red's second choice f_2, and an interaction between f_1 and f_2. Using split-half cross-validation prevented us from adding further interactions, which led to singular design matrices.

Experiment

Methods

Participants We recruited 100 US participants on Prolific.co (53 male, 46 female, 1 non-binary), aged 42 ± 13 years old. Participants were paid a flat rate of $12 per hour and the study lasted 14 ± 3.76 minutes.

Design The study was a 2×2×4 within-subject design. First, we varied whether Red and Green had competitive (i.e. both prefer apples) or complementary (i.e. Red prefers oranges while Green prefers apples) preferences. Second, we varied the fruits available to choose from in scenes 2 and 3: either one apple and one orange, or two apples and one orange. Finally, we varied the choices that Red made when Purple was present and when they were absent, yielding four combinations: apple, apple; apple, orange; orange, apple and orange, orange. Overall, combining these conditions yielded 16 stories and spanned cases where Red's personal, social and presentational goals align or conflict. The experiment was built using JsPsych (De Leeuw et al., 2023). The study was preregistered on OSF. The preregistration along with the data and analysis scripts are available at https://github.com/Vbtesh/cogsci2026_presentational_goals. The model variants aimed at improving the fit for δ were not preregistered and emerged from the observation that the joint-posterior marginal struggled with recovering δ.

Procedure The study begins with participants seeing instructions about the experiment. Participants are presented with three stories for which they have to answer basic comprehension questions such as: what does Red/Green prefer? Does Purple know what Red/Green prefers? What did Red pick in scene 2 or scene 3? Participants have to answer all of those questions correctly in order to proceed. The main section of the experiment involves two blocks presented in a random order. The first block consists of every story in which the preferences of Red and Green are in competition, that is both Red and Green prefer the apple. We then present, in a random order, every combination of choices Red can make for scene 2 and scene 3. For example, apple when Purple is absent then apple when Purple is present, or apple when Purple is absent then orange when Purple is present. As each set of choices can be applied to stories with one apple and one orange or stories with two apples and one orange, this block is 8 stories long. Following the first block, we let participants know that in the next set of stories, Red will prefer Oranges while Green will prefer Apples. We then repeat the same

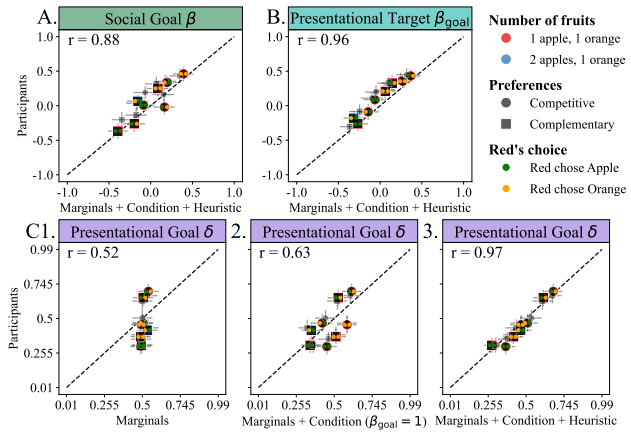

Figure 3: **Correlations between participants' judgments and model marginals from the joint posterior**. (A) social goal β (how does Red feel about Green?), (B) presentational target β_{goal} (what does Red want Purple to think?), and (C) presentational goal δ (how much does Red care about what Purple thinks of Red?). Panels C2 and C3 show predictions for δ from the model variants that additionally condition on $\beta_{\text{goal}} = 1$ and also add the heuristic. Error bars are bootstrapped 95% CIs.

set of 8 possible combinations of choices and available fruits. After seeing all 16 stories, the experiment ends with a short demographics questionnaire.

After each story, participants were asked three questions, answered on continuous sliders, which we propose map onto three key parameters of the model (see Figure 1). We counterbalanced block order, competitive or complementary preferences first, the order of the stories within each block, and the order of scenes 2 and 3 in each story.

Results and Discussion

Fitting procedure The feature-based model was fitted using the `MixedLM` method of the `statsmodels` Python package. All other variants were built in `Memo` (Chandra et al., 2025) and fitted with the `Optax` Python package (Bradbury et al., 2018). We used gradient descent from random initialisation of parameters using a cross-entropy loss on the empirical distributions, that is we jointly minimised the Kullback-Leibler (KL) divergence between participants' responses and the marginal posteriors. We cross-validated each model using 100 folds of half-splits: we fitted the model to 8 randomly selected stories and predicted the other 8 for 100 sets of 8 stories.

Systematic intuitions about presentational goals Figure 2 shows participants' judgments along with predictions from the model using both marginals and the heuristic for δ. We selected 8 stories highlighting the richness and diversity of participants' inferences. In stories B and E, Red's social goals are relatively clear, as Red consistently acts opposite to their personal preference in the latter two scenes.Story C illustrates

Table 1: **Model comparison**. Median and [5th, 95th] percentiles of correlations (higher is better) and KL divergences (lower is better) for 100 folds of split-half cross-validation for each model. The non-feature-based models differ only in how they estimate δ.

Pearson's r ↑	β	δ	β_{goal}
Feature-based	**0.93** [0.60, 0.98]	**0.97** [0.87, 0.99]	0.93 [0.66, 0.99]
Marginals	0.86 [0.69, 0.95]	0.43 [-0.30, 0.80]	0.95 [0.87, 0.98]
Marginals + Cond. ($\beta_{\text{goal}} = 1$)	0.87 [0.69, 0.96]	0.62 [0.30, 0.82]	**0.96** [0.93, 0.98]
Marginals + Cond. + Heu.	0.87 [0.74, 0.97]	0.95 [0.87, 0.98]	**0.96** [0.91, 0.99]
KL divergence[1] ↓			
Feature-based	1.23 [1.03, 1.46]	1.52 [1.37, 1.69]	1.14 [0.94, 1.31]
Marginals	**0.81** [0.68, 0.98]	0.98 [0.86, 1.17]	0.81 [0.68, 0.98]
Marginals + Cond. ($\beta_{\text{goal}} = 1$)	0.82 [0.67, 0.97]	0.95 [0.75, 1.37]	**0.80** [0.66, 0.94]
Marginals + Cond. + Heu.	**0.81** [0.66, 0.95]	**0.73** [0.60, 0.92]	**0.80** [0.66, 0.93]

[1]Standardised KL divergence with the uniform distribution as reference: lower than 1 is closer to participants than the uniform, higher is further.

a canonical case of Red yielding their preference to Green in the third scene, when Purple is present, but not in the second scene. Both participants and our model predict that Red has a strong, positive presentational goal (wants to be well perceived by Purple), but both have more uncertainty about Red's social goal. However, note that our model still considers that the story gives slightly more weight to Red liking Green than disliking them. Indeed, even if Red has a clear presentational goal, it is not enough to explain away the possibility that they do like Green: after all, they chose to pick an orange when they prefer apples. Participants seem slightly harsher than our model in this case.

The flexibility of presentational targets It is also possible to have negatively valenced presentational goals, such as wanting to hide a friendship from others or be perceived as adversarial. Our model lets agents have presentational targets over any value of β, including when β_{goal} is negative. For example, in story D, Red lets Green take the apple when Purple is absent, but then acts selfishly when Purple is present. Like participants, our model infers that Red truly likes Green, but wants Purple to think that they do not. This may be akin to two childhood friends where one of them is not well liked at school, leading the other to pretend like they are not close in front of other children. Notice the bimodality in the distribution over δ, which may stem from different interpretations of presentational goals. One interpretation ties presentational goals to wanting to be well seen by Purple, which should be low here given that $\beta_{\text{goal}} < 1$. An alternative interpretation is that Red does not actually care what Purple thinks given that their behaviour changed in Purple's presence. Our model is able to account for both interpretations.

Table 2: **Parameter fits** for each model variant.

Model	τ_{obs}	τ	θ	σ_{goal}	T^{-1}	α	σ_{heu}
Marginals	3.50	1.75	0.30	0.47	2.05		
Marginals + Cond. ($\beta_{\text{goal}} = 1$)	7.46	0.41	0.11	0.51	2.07		
Marginals + Cond. + Heu.	5.98	0.47	0.15	0.50	2.30	0.33	0.31

Comparison with feature-based model We compare cross-validated models using Pearson's correlation between mean judgments and predictions, as well as KL divergence to measure how well models can capture more complex features of participants' responses (see Table 1). Overall, we see that while the feature-based model captures the mean parameters for each story for each judgment, our model does so just as well with fewer parameters. However, KL divergences between the feature model predictions and participants' judgments are strictly worse than a uniform distribution. This is because the feature model can only predict approximately Gaussian distributions centred on each story's mean prediction. However, as Figure 2 shows, participants' responses are more nuanced. In comparison, our model is strictly better than a uniform distribution at capturing patterns of responses (see full distributions in Figure 2).

Comparing model variants to interpret the δ result The pattern in Figure 3 where marginals from the joint show strong recovery of β ($r = 0.88$) and β_{goal} ($r = 0.96$) but poor recovery of δ unless we condition on $\beta_{\text{goal}} = 1$ or add the behaviour change heuristic warrants specific discussion. We highlight three points.

First, the strongest validation of our model lies in our ability to recover the presentational *target* β_{goal} ($r = 0.96$). Participants' inferences about what Red wants Purple to believe highlight the need for the added layer of recursion that presentational goals necessitate. It requires that the observer simulates Red simulating Purple, and our model captures this process based on evidence from only two scenes. Second, the gap from the simple marginal over δ ($r = 0.52$) to conditioning on $\beta_{\text{goal}} = 1$ ($r = 0.63$) is informative. This suggests that some participants interpreted "how much does Red care about what Purple thinks" as "how much does Red want Purple to think they are prosocial". Future work could reverse the order of presentational questions and ask "what does Red want Purple to think?" followed by "how much does Red care that Purple thinks that?". Third, the final gain from using the behaviour change heuristic ($r = 0.97$) with mixing weight $\alpha = 0.33$ allows us to capture most of the remaining variance by accommodating bimodality in the distribution of responses (see stories B or D in Figure 2). One natural interpretation is that participants marginalise over latent variables which our model considers fixed, such as hierarchical and power structures (Davis et al., 2026; Gershman & Cikara, 2020).

Taken together, these results suggest that while δ, the weight on presentational goals, is harder to infer from sparse behavioural evidence, our model's ability to recover β and most notably the presentational target β_{goal}, which requires an added layer of recursive reasoning, strongly supports the inverse planning account.

General Discussion

The present work introduced a novel paradigm and computational model to study how people jointly infer the different goals that others may have. Our results demonstrate that people can systematically and jointly infer an agent's social goal (how much they like or dislike another agent), presentational goal (how much they value what another agent thinks), and presentational target (what they want another agent to think about them). Our approach captures a wide range of social interactions where these goals may be misaligned, such as when an agent dislikes another but wants to appear prosocial (Figure 2C), or when an agent likes another but wants to hide this fact (Figure 2D). With minimal assumptions, our model adapts to each of those situations and matches participants' intuitions, providing further support for inverse planning as a meaningful model of how people navigate the social world (Baker et al., 2009, 2017; Jara-Ettinger et al., 2016).

Furthermore, our findings speak to the computational architecture of theory of mind and what human social utilities are *made of*. First, formalising presentational goals as information-theoretic utilities reveals that what humans consider to be costs and rewards in social interactions goes beyond material outcomes. Instead, they can rely on computations over changes in others' beliefs which may matter more than material payoffs. Second, while classic false-belief tasks, such as the Sally-Anne task (Wimmer & Perner, 1983), test first-order theory of mind, presentational goals require representing what others believe about our social preferences. The fact that participants were systematic in their inferences and were well captured by a rational Bayesian model suggests that human social cognition generally handles this recursion.

Finally, while positively valenced presentational goals may be the norm — we generally want others to think we are *prosocial* — the cognitive process which allows us to infer this goal is not limited to these cases. Our findings suggest that managing impressions rests on a valence-agnostic engine which flexibly adapts to less common social interactions. This valence-agnostic property may be relevant in linking presentational goals to higher order inferences about relational structures. People are remarkably adept at inferring social structures, hierarchies, and group memberships, yet formal accounts of the intuitive theories underlying these inferences are still nascent and often treat sociological knowledge as a primitive (Davis et al., 2026; Gershman & Cikara, 2020). Presentational and social goals offer an alternative view which builds those from lower level inferences about individuals' revealed preferences. If A cares what B thinks but B does not care what A thinks, the asymmetry maps onto subordination or mentorship. Alternatively, if presentational concern is mutual between A and B but absent towards C, the structure resembles friendship cliques. Valence matters for managing the complexity: shared negative presentational targets towards an outside party can reveal a hidden alliance while shared positive targets may indicate deference towards a common authority. This approach would connect inferences about individual preferences to higher-order constructs without having to posit the existence of a separate grammar (Jara-Ettinger & Dunham, 2025).

Acknowledgements

We thank Kartik Chandra for valuable discussions and feedback. VB was supported by a Bogue Fellowship and an UCL Experimental Psychology Demonstratorship. TG was supported by grants from the Stanford Institute for Human-Centered Artificial Intelligence (HAI) and from the Cooperative AI Foundation.

References

Baker, C. L., Goodman, N. D., & Tenenbaum, J. B. (2008). Theory-based social goal inference. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*.

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.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Battigalli, P., & Dufwenberg, M. (2022). Belief-Dependent Motivations and Psychological Game Theory. *Journal of Economic Literature*, *60*(3), 833–82.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., & Zhang, Q. (2018). JAX: Composable transformations of Python+NumPy programs.

Btesh, V., Lagnado, D., & Gerstenberg, T. (2025). Taking others for granted: Balancing personal and presentational goals in action selection. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Chandra, K., Chen, T., Li, T.-M., Ragan-Kelley, J., & Tenenbaum, J. (2024). Cooperative Explanation as Rational Communication. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Chandra, K., Chen, T., Tenenbaum, J. B., & Ragan-Kelley, J. (2025). A Domain-Specific Probabilistic Programming Language for Reasoning about Reasoning (Or: A Memo on memo). *Proc. ACM Program. Lang.*, *9*.

Davis, I., Jara-Ettinger, J., & Dunham, Y. (2026). Inferring the internal structure of groups through the integration of statistical learning and causal reasoning. *Nature Communications*, *17*(1).

De Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8*(85), 5351.

Degen, J. (2023). The Rational Speech Act Framework. *Annual Review of Linguistics*, *9*, 519–540.

Gershman, S. J., & Cikara, M. (2020). Social-Structure Learning. *Current Directions in Psychological Science*, *29*(5), 460–466.

Goffman, E. (1959). *The presentation of self in everyday life*. Doubleday.

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910.

Houlihan, S. D., Kleiman-Weiner, M., Hewitt, L. B., Tenenbaum, J. B., & Saxe, R. (2023). 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.

Jara-Ettinger, J., & Dunham, Y. (2025). The Institutional Stance. *Behavioral and Brain Sciences*, 1–62.

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

Kleiman-Weiner, M., Shaw, A., & Tenenbaum, J. B. (2017). Constructing Social Preferences From Anticipated Judgments: When Impartial Inequity is Fair and Why? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *39*.

Netanyahu, A., Shu, T., Katz, B., Barbu, A., & Tenenbaum, J. B. (2021). Phase: Physically-grounded abstract social events for machine social perception. *Proceedings of the aaai conference on artificial intelligence*, *35*(1), 845–853.

Powell, L. J. (2022). Adopted Utility Calculus: Origins of a Concept of Social Affiliation. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *17*(5), 1215–1233.

Roberts, G. (2020). Honest signaling of cooperative intentions. *Behavioral Ecology*, *31*(4), 922–932.

Shu, T., Kryven, M., Ullman, T. D., & Tenenbaum, J. B. (2020). Adventures in Flatland: Perceiving Social Interactions Under Physical Dynamics. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *42*.

Tejwani, R., Kuo, Y.-L., Shu, T., Katz, B., & Barbu, A. (2022, November). Social Interactions as Recursive MDPs. In A. Faust, D. Hsu, & G. Neumann (Eds.), *Proceedings of the 5th Conference on Robot Learning* (pp. 949–958, Vol. 164). PMLR.

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

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Wu, S. A., Sridhar, S., & Gerstenberg, T. (2023). A computational model of responsibility judgments from counterfactual simulations and intention inferences. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*, 3375–3381.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020). Polite Speech Emerges From Competing Social Goals. *Open Mind: Discoveries in Cognitive Science*, *4*, 71–87.

Zhao, X., Zhao, X., Gweon, H., & Kushnir, T. (2021). Leaving a Choice for Others: Children's Evaluations of Considerate, Socially-Mindful Actions. *Child Development*, *92*(4), 1238–1253.