Taking others for granted: Balancing personal and presentational goals in action selection

Victor J. Btesh[1*], **Tobias Gerstenberg**[2], **David A. Lagnado**[1]
[1]University College London, [2]Stanford University [*]victor.btesh.19@ucl.ac.uk

Abstract

This study investigates how individuals balance personal and presentational goals, i.e. how they wish to be perceived, in social interactions where those are conflicting. Using a computational model framing presentational goals as minimising the divergence between the perceived and desired belief state of their partner, e.g. how much their partner trusts them versus how much they would want to be trusted, we predict complex decision-making patterns which cannot arise from solely focusing on maximising their partner's utility. In accordance with our model, participants tended to forego signaling good intentions and prioritised their own goals when they perceived their partner to trust them. As further predicted, participants were less concerned about how they were perceived and acted in their own interest more often when their partner was unlikely to change their mind. We discuss implications and potential extensions to this framework.

Keywords: presentational goals; theory of mind; Bayesian models; social cognition

Introduction

Who has not delayed meeting a friend at the last minute to finish a trivial task at home or to spend a little more time with someone else? This puzzling tendency raises the question: why are we sometimes careless about those we care about and who value us? In this paper, we want to demonstrate that this pattern of behaviour is consistent with a relatively simple mathematical formulation for how people manage their material and social goals. The dialectics of pursuing goals which are misaligned with how we would want to be perceived are a fundamental aspect of life in a society. Managing those constant micro-conflicts requires that we can both evaluate the relative importance of a goal and weigh it against a potential cost in reputation. This necessitates holding a representation of how our choices will affect others' beliefs about us and our goals and intentions. Put differently, we ought to be balancing local material gains with the changes in others' beliefs about us that pursuing those gains would cause.

Inverse planning and intervening on beliefs Inverse planning uses probabilistic reasoning to infer an individual's beliefs, goals, and preferences by observing their actions, working backward from behavior to understand their intentions (Baker, Saxe, & Tenenbaum, 2009; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). Social Markov Decision Processes (MDPs) formalise how agents resolve conflicts between material and social goals through recursive inference of others' value functions, integrating these into their own (Tejwani, Kuo, Shu, Katz, & Barbu, 2022). Agents can act strategically to signal specific intentions, knowing that their actions will be used by others to infer their goals (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017). Research has shown how inferring others' beliefs and goals also allows agents to intervene on mental states. For example, Houlihan, Kleiman-Weiner, Hewitt, Tenenbaum, and Saxe (2023) demonstrated that inverse planning, event appraisals, and inferred social preferences can predict nuanced emotions in high-stakes social dilemmas. Recently, storytelling and explanations have been framed as interventions on beliefs, where narrating events guides listeners to desired conclusions (Chandra, Chen, Li, Ragan-Kelley, & Tenenbaum, 2024; Chandra, Li, Tenenbaum, & Ragan-Kelley, 2024). People adeptly predict how their actions influence others' beliefs and use these predictions to guide behavior (Ho, Saxe, & Cushman, 2022; Oey, Schachner, & Vul, 2023).

These models provide provide great insights but frame interactions as driven by expected rewards, representing others only when agents care about their goals (Cushman, 2024; Jara-Ettinger et al., 2016; S. A. Wu, Sridhar, & Gerstenberg, 2023). Such agents only help or hinder to maximize their own gains or due to intrinsic motives (Ullman et al., 2009). However, even without clear intentions, people may still choose to act amicably or deceitfully to maintain their reputation, prioritizing others' perceptions over their actual outcomes.

Signaling and prosocial behaviour Prosocial behavior, actions which benefit others (Y. E. Wu & Hong, 2022), is studied in economics as an intrinsic preference for fairer outcomes (Fehr & Schmidt, 1999) or as an aversion to negative emotions like guilt or frustration (Aina, Battigalli, & Gamba, 2020; Battigalli & Dufwenberg, 2007, 2022). It can stem from empathy, reciprocity (Saulin, Horn, Lotze, Kaiser, & Hein, 2022), or to build and maintain trust (Bicchieri, Xiao, & Muldoon, 2011; Jordan, Hoffman, Nowak, & Rand, 2016). Jordan et al. (2016) showed that uncalculating cooperation in economic games signals trustworthiness when decision processes are observable. We propose that such behavior, while seemingly ignoring reward-based motives, may conceal computations about partners' beliefs.

Presentational goals We propose that prosocial, but also adversarial behaviour can be modelled not only as an intrinsic preference for equitable outcomes or to avoid negative emotions but as a means of influencing the beliefs others have about us. If humans are equipped with the ability to infer others' intentions from their actions and use these inferences to affect others' beliefs, then they may also hold preferences about beliefs that others have of them (Gweon, 2021). We refer to this specific type of social goal as presentational goals (see Yoon, Tessler, Goodman, & Frank, 2020). Crucially, the target for such goals ceases to be the actual utility of other agents but some function of what we want them to think our intentions are.

Computational model

We propose a computational model of how people manage situations where material goals compete with presentational goals. To do this, we propose an adapted version of the dictator game (Camerer & Thaler, 1995; Guala & Mittone, 2010). Suppose two agents, Blue B and Green G, are interacting. Specifically, B can choose to take an action a from a set A which yields reward $R_B(a)$ for B and $R_G(a)$ for G, known to both agents. Crucially, G can only observe what B does: they cannot act. We assume that each agent evaluates their own gains as well as the other's gains using the following utility function. For agent B:

$$U_B(a) = \alpha \cdot R_B(a) + \beta \cdot R_G(a) \tag{1}$$

where α represents self-interest (how much B cares about their own reward) and β prosociality (how much they care about the other agent's gain).

The key contribution of this paper is to suppose that social agents attempt to influence the belief others have of the values of their parameters α and β (see also Yoon et al., 2020). That is, they assume that other agents attempt to infer the value of α and β from their behaviour using Bayesian inference. Our approach is inspired by social MDPs (Tejwani et al., 2022) but instead of only using the outputs for optimising their gains and plan around other agents, they also use it to influence each other's beliefs. Here, we consider that B wants to influence G's estimate of B's prosocial parameter β using a presentational term $\text{Pres}(a, \beta_G)$ where β_G is G's estimate of β. The full utility function is thus

$$U_B(a) = \alpha \cdot R_B(a) + \beta \cdot R_G(a) + \delta \cdot \text{Pres}(a, \beta_G), \tag{2}$$

where δ represents the contribution of the presentational goals of B in their decision making. The key challenge is thus to model $\text{Pres}(a, \beta_G)$. We propose that agent B attempts to bring a posterior distribution $p(\beta_G|a)$, i.e. what G would believe about them after taking action a, in line with a distribution $\rho(\beta_G)$ representing their preference about what $p(\beta_G|a)$ ought to be. As such, we write the social target as

$$\text{Pres}(a, \beta_G) = -D_{KL}[p(\beta_G|a)||\rho(\beta_G)], \tag{3}$$

which implies that B will be trying to find an action which minimises the Kullback-Leibler divergence between the estimated posterior $p(\beta_G|a)$ conditional on taking action a and their preference $\rho(\beta_G)$. From there, agent B chooses an action a from the action set A by sampling from a softmax distribution, i.e. $p(a) = \sigma(U_B(a))$.

Inference In order to compute $\text{Pres}(a, \beta_G)$, B must be able to anticipate how taking action a would influence G's estimate β_G of their prosocial parameter β. As such, agent B must be able to find or estimate $p(\beta_G|a)$. For simpliciy, B assumes that G is not inferring their presentational goals parameter δ. We further assume a mean-field distribution for the joint distribution over parameters, namely that all parameters are mutually independent: $p(\alpha_G, \beta_G) = p(\alpha_G)p(\beta_G)$. Focusing on β_G, it is straightforward to show that

$$p(\beta_G|a) \propto \exp(\beta_G R_G(a))p(\beta_G) \tag{4}$$

Intuitively, this suggests that to infer what other people would believe their prosociality to be given some action a, agents use an average of their utility given the action weighted by their current estimate of what others think of them.

Predictions All predictions hinge on the fact that the behaviour of an agent with presentational goals is dependent on the current beliefs of other agents. Suppose an agent B is not prosocial but wishes to be perceived as such, therefore $\beta = 0$ and $\delta > 0$. Explicitly writing $\text{Pres}(a, \beta_G)$ as $-D_{KL}[p(\beta_G|a)||\rho(\beta_G)]$, their utility is

$$U_B(a) = \alpha \cdot R_B(a) - \delta \cdot D_{KL}[p(\beta_G|a)||\rho(\beta_G)]$$

As $p(\beta_G|a) \to \rho(\beta_G)$, then $D_{KL}[p(\beta_G|a)||\rho(\beta_G)] \to 0$ and so the contribution of the presentational terms decreases.

Hypothesis 1: When presentational goals are achieved or close to being achieved, agent B will tend to forego signaling prosociality.

The contribution of $D_{KL}[p(\beta_G|a)||\rho(\beta_G)]$ in selecting action a_i rather than action a_j depends on the magnitude of their respective changes in $p(\beta_G|a)$. That is, if all actions a have little influence (i.e. if $p(\beta_G|a_i) \approx p(\beta_G|a_j)$ for all a_i, a_j), then $D_{KL}[p(\beta_G|a)||\rho(\beta_G)]$ becomes almost constant and hence no longer discriminates between actions. A trivial case occurs when the beliefs of G are independent from B's actions. A more interesting case is when G is so sure about B's intentions that any action B takes cannot meaningfully change their mind. Formally, when the entropy of their distribution over β_G gets very low: $H(p(\beta_G)) \to 0$. A third case occurs when G is unable to observe B's actions or their consequences, or at least when B believes this to be case. Then, B's actions cannot change G's beliefs.

Hypothesis 2: When B believes G to be certain in their beliefs, i.e. when those have low entropy, B should value presentational goals less if those involve convincing them of a specific hypothesis, e.g. convincing them that they are prosocial.

Hypothesis 3: Agent B should pursue their own interests more when they are convinced that they cannot be observed.

Alternative models As points of comparison, we fit lesioned variants of our model. One with no presentational

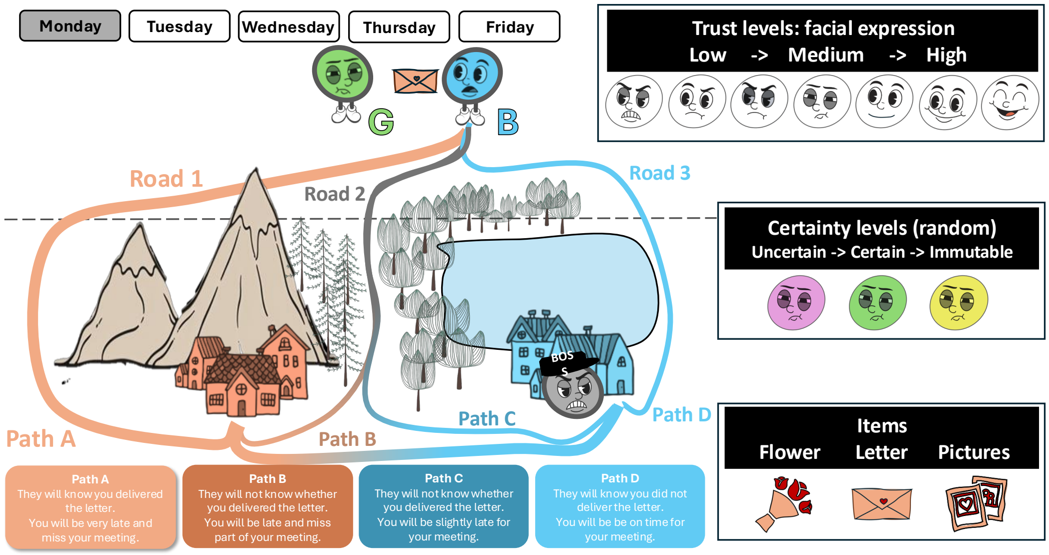
Monday
Tuesday
Wednesday
Thursday
Friday
G
B
Trust levels: facial expression
Low -> Medium -> High
Road 1
Road 2
Road 3
Certainty levels (random)
Uncertain -> Certain -> Immutable
BOSS
Path A
Path B
Path C
Path D
Items
Flower
Letter
Pictures
Path A
They will know you delivered the letter.
You will be very late and miss your meeting.
Path B
They will not know whether you delivered the letter.
You will be late and miss part of your meeting.
Path C
They will not know whether you delivered the letter.
You will be slightly late for your meeting.
Path D
They will know you did not deliver the letter.
You will be be on time for your meeting.

Figure 1: Materials used for the experiment. Each day, participants could choose a path from the four options. Path A signaled to G that the item would be delivered. Path B and C would give no signal but the former would deliver the item while the latter would not. Finally, path D would not deliver the item and signal it to G. Trust level was mapped onto facial expression as shown above, certainty was assigned randomly to a character colour/item pair which stay constant throughout the experiment.

goals, i.e. $\delta = 0$, and one with no social goals, $\beta = 0$.

Methods

Materials To test our hypotheses, we introduce a version of the dictator game inspired by Chandra, Li, Tenenbaum, and Ragan-Kelley (2023), where an agent B, played by participants, is tasked by a *green*, *violet* or *yellow* agent, which we will always call G, played by a bot, to deliver an item to their friend in the Red village (see Figure 1). B must choose between three roads to reach Blue village for a work meeting. Road 1 is the longest and will anger their boss but involves passing through Red village and thus signals to G that the item will be delivered. Road 2 is ambiguous, potentially allowing B to deliver G's item with less time cost but still angering their boss. Road 3 is the shortest, ensures B is on time, but clearly reveals that G's item will not be delivered. Participants were explicitly told G would not know whether the item was delivered, making Road 2 potentially more desirable despite the trade-offs. Consequently, while R can only see which road B takes, i.e. 1, 2 or 3, B has four options (see Figure 1). This paradigm was built in Otree (D. L. Chen, Schonger, & Wickens, 2016).

Representing agent G Given that participants were playing as B and we needed to have a controllable behaviour for G, we used a bot. The bot updated their beliefs $p(\beta_G|a)$ based on their prior, of which we manipulated the entropy, and on actions taken by participants. In order to perform those updates, we had to define a formal reward structure for the task but did not disclose it to participants. Paths A, B, C and D corresponded to a reward of 0, 1, 2 and 3 for B, representing how satisfied their boss would be based on when they arrive. For agent G, the payoffs were respectively 4, 4, 0, 0 corresponding to either delivering or not delivering the letter. Crucially, as G could only tell which road B takes, i.e. 1, 2 or 3, their perceived payoffs had the two middle paths averaged. So the payoffs for B were 0, 1.5, 3 and 4, 2, 0 for G. G simply computed $p(\beta_G|a) \propto \exp(\beta_G R_G(a))p(\beta_G)$, where β_G could take values 1 for prosocial, -1 for adversarial and 0 for neither. The mean of the posterior distribution was then mapped onto 7 faces representing the following emotions *very unhappy, unhappy, slightly unhappy, unsure, slightly happy, happy, very happy* and displayed to participants (see Figure 1). Participants were told that each facial expression represented a level of trust for G: how much they believed that B cared about actually delivering the item.

Design We used a 3×3 within subject design with $9 + 1$ blocks of trials. The first condition, targeted at *hypothesis 1*, used facial expressions to manipulate the starting belief state of G. We had three cases where for β_G in $[-1, 0, 1]$, trust was *high* with $p(\beta_G) = [0.1, 0.2, 0.7]$, *moderate* with $p(\beta_G) = [0.15, 0.7, 0.15]$ and *low* with $p(\beta_G) = [0.7, 0.2, 0.1]$. The second condition, targeted at *hypothesis 2*, used different character colours and item pairs to test whether participants were sensitive to how certain of their beliefs G was. We set the entropy of G's prior $p(\beta_G)$ by passing it through a softmax with inverse temperature parameter τ^{-1} to lead G to be *uncertain* with $\tau_u^{-1} = 1$, *certain* with $\tau_c^{-1} = 10$ and *immutable* with $\tau_i^{-1} = 100$. The latter being an extreme case where G simply never updated their beliefs irrespective of B's actions. To test *hypothesis 3*, we added an additional block where G was absent: that is where they will be not observe B's actions at all. The game was organised in weeks, whereby participants were told that they would visit G on some business day of the week and that G's mental states would be reset each weekend. This way, we could change G's beliefs between weeks and test the different combinations of conditions. Fur-

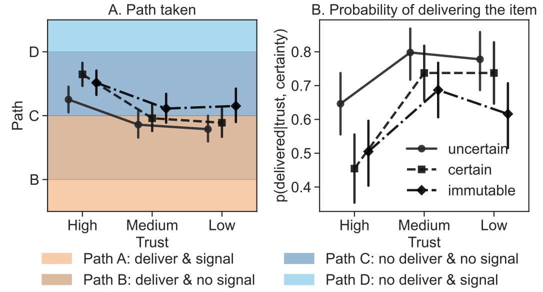

Figure 2: Path taken (A.) and probability of delivering the item (B.) given trust and certainty levels.

thermore, having participants interact with G for more than a week was a way to train them on G's certainty level, i.e. entropy of their prior. For participants to learn G's certainty, we explicitly stated that G was the same person throughout the game.

Participants We recruited 99 British and American participants ($M = 51$, $F = 45$, $NB = 2$ and 1 preferred not to say) aged 34 ± 13 from Prolific. The task lasted about 30 minutes and participants were paid a £4.5 flat fee for taking part in the study. We did not provide any bonus to let people intuitively weigh each choice. To control for inattentive participants, we positioned attention checks before randomly chosen trials. We excluded one participant for failing more than three attention checks.

Procedure Following a consent form, participants were explicitly told that all their judgments should be intuitive and that their payment would not be affected by their decisions. They then followed an introduction to the game and the characters, going through the materials described above. To check their understanding, they had to complete a validation quiz before being able to proceed. The experiment was split in two parts. Part 1 involved visiting G every business day and was comprised of 10 blocks of 5 trials. Each block had participant interact with a different version of G corresponding to a certainty and trust level pair. For instance, if *uncertain* was associated with the green character and the letter, then in the *uncertain* - *low trust* block, a participant would be asked by a green character with an initially unhappy face to deliver their letter. Critically, after every trial within each week, participants saw how their path choice affected G's state of mind, allowing them to learn and plan subsequent decisions. Part 2 repeated the certainty-trust pairs of part one but only for a single day and had thus 10 blocks each with a single trial. Instead of displaying G's updated state, we asked participants to predict how they thought their actions would change G's state of mind. Following these, participants were asked to report how much they weighed being on time, and how much they valued pleasing G. They also provided a brief qualitative summary of how they approached situations given G's initial state. Finally, participants completed a brief demographics questionnaire.

Results

For all analyses below, we focused on the 10 blocks in part 2, which were one-shot judgements about which path to take under different certainty and trust conditions. The data, analyses scripts and modelling outputs for Stan can be found here: https://anonymous.4open.science/r/taking_others_for_granted-0B4A/.

Behavioural results

To establish whether the choice of path, i.e. how happy will B's boss be, depended on the trust level of agent G and on their certainty level, i.e. how likely they are to change their mind, we run a mixed-effect regression predicting how late participants were to work and consequently how angry their boss was from the character's trust and certainty levels (Singmann et al., 2024) (see Figure 2). The reaction of the boss was measured on a scale 1 to 4, corresponding respectively to the shortest and longest paths. The trust of G had three levels: *high*, *medium* and *low*. Finally, certainty also had three levels: *uncertain*, *certain* and *immutable* corresponding in the task to the different character colours. We find a fixed effect of certainty ($F(2, 104.81) = 7.09, p < .01$) and trust ($F(2, 102.41) = 20.1, p < .001$) but no significant interaction. Trials where trust was *high* ($\mu = 2.53 \pm .08$) had participants be significantly more timely than those where trust was *medium* ($\mu = 3.02 \pm .08$), $t(98) = -6.04, p < .001$, or those when it was *low* ($\mu = 3.06 \pm .08$), $t(98) = -5.0, p < .001$. In accordance with *hypothesis 1*, this suggests that participants were more likely to favour their boss's reaction and be on time at work when characters already felt positively about them. Along the certainty dimension, we observe a significant difference between trials with the *uncertain* character ($\mu = 3.03 \pm .07$) versus those with the *certain* ($\mu = 2.84 \pm .07$), $t(98) = 2.7, p = .022$ or the *immutable* ($\mu = 2.74 \pm .07$), $t(98) = 3.58, p < .001$) characters. Consistent with *hypothesis 2*, this suggests that when the character was less likely to change their mind, participants tended to focus more on arriving early to please their boss.

We further wish to establish when participants were more likely to not deliver the item, i.e. taking advantage of the trust and certainty of G to not fulfill their request. We ran another mixed effect model, predicting the likelihood of actually delivering the item from the same predictors. We find a similar pattern of results with a stronger fixed effect of certainty ($F(2, 98) = 19.93, p < .001$) and of trust ($F(2, 98) = 27.7, p < .001$) with no significant interaction. The probability to deliver the item went from an estimated 79% with the *uncertain* down to 69% with the *certain* character ($z = 2.784, p = .015$) and 63% with the *immutable* character ($z = 4.1, p < .001$). This suggests that *uncertain* characters, those with a fluctuating mood, were more likely to get their item delivered than the other two. For trust, the probability of delivering the item went from an estimated 50% for the *high trust* trials up to 79% for *neutral* ($z = -5.8, p < .001$) and 77% in the *low* trust trials ($z = -5.0, p < .001$). This

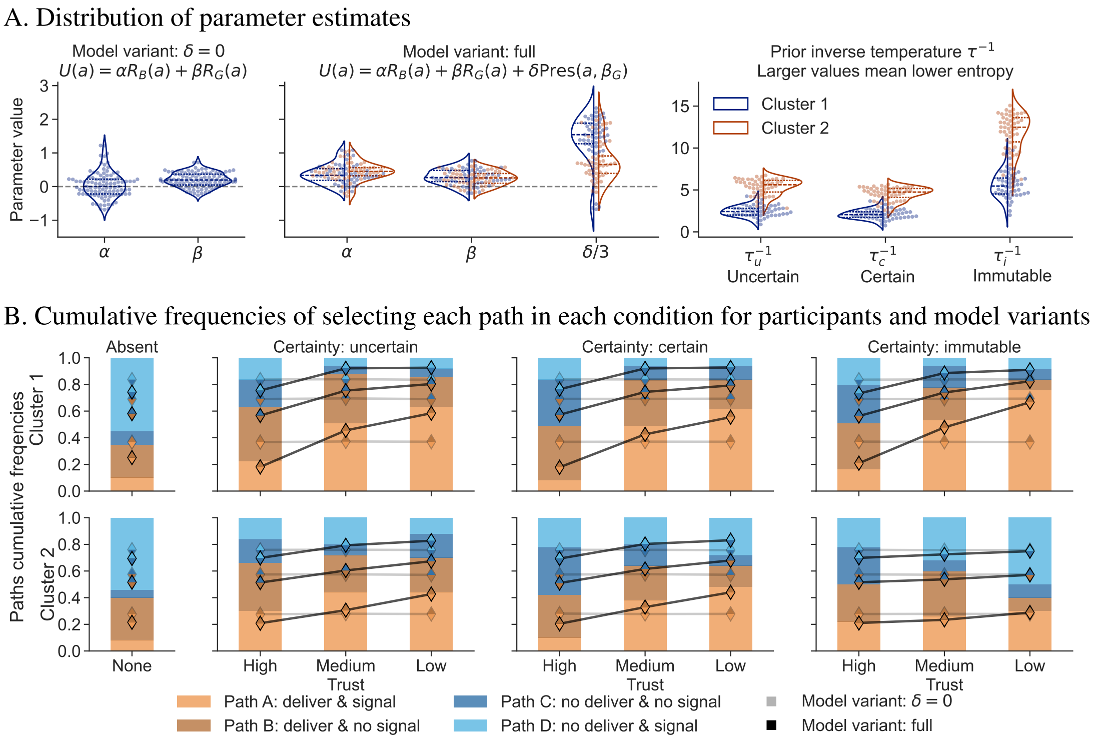

Figure 3: A. Distribution of parameter estimates for the $\delta = 0$ and full models for each recovered cluster. δ has been rescaled to be in a comparable unit to α and β as its unit in KL divergence made the estimates much larger. The last plot provides the inverse temperatures controlling the entropy of the prior over the character's beliefs. Larger values mean lower entropy and thus more certainty. B. Participants (stacked bars) and model variants (points) cumulative frequencies for each path for each cluster (rows) in all certainties (columns) and all trust conditions (x-axis)

shows that when characters were trusting, participants were much less likely to fulfill their request.

Finally, in accordance with *hypothesis 3* we verified that without the presence of G, participants were significantly more likely to focus on arriving on time and making their boss happy rather than delivering the item ($F(2,98) = 75.23, p < .001$)). Consistent with Bicchieri et al. (2011), participants were more timely when G was absent ($\mu = 1.92 \pm .11$) than where they were present ($\mu = 2.87 \pm .06$): $t(98) = 8.673, p < .001$.

Modelling

Table 1: Model fits for all model variants and per cluster. r^2 were computed from 400 bootstrapped samples

Model	Fit(r^2)	cluster 1	cluster 2
$\delta = 0$	.25	.31	.07
$\beta = 0$	.54	.68	.16
full	**.62**	**.76**	**.22**

To provide a more complete account of participants' behaviour, we fitted variants of the model to the data in the 10 testing blocks using Stan. Model variants had two families of parameters. First, the weights of the utility function: α, β and δ (see Equation 2). Second, a set of inverse temperatures controlling the entropy of the estimated prior of the character, i.e. their level of certainty: τ_u^{-1}, τ_c^{-1} and τ_i^{-1}. For stability, we fitted $\log \tau^{-1}$. Each parameter had a group and participant level term which combined linearly. We fit a model with no presentational goals, i.e. one where $\delta = 0$, a model with no social goals, i.e. $\beta = 0$ and and a full model with social and presentational goals. An apparent bimodality in the distribution of the parameters for the full model motivated us to use a Gaussian mixture model on parameter values to observe whether we could recover two latent groups of participants. The clusters were evenly distributed with 49 participants in cluster 1 and 50 in cluster 2. Table 1 summarises model fits. We see that the full model performed better overall ($r^2 = .62$). Importantly, across all variants, the model performed much better for participants in cluster 1 ($r^2 = .76$) than cluster 2 ($r^2 = .22$). Figure 3a shows the distribution of parameter estimates for each cluster of participants and Figure 3b shows participants' judgments and the predictions for variant with no presentational goals and the full model.

We can see from Figure 3b that the presentational δ term was necessary for capturing the effect of *trust* on behaviour: only the variant with a nonzero δ parameter captures the fact that participants tend to prioritise trusting G characters less than others. Note that parameter estimates for the full model show that participants tend to behave as if they had baseline level of pro sociality, i.e. $\beta > 0$. We can see evidence of this by the large frequency of path B, i.e. the path which delivers the item but does not signal it. Crucially, we can see qualitatively different behaviours between the two clusters.

Two strategies Participants in cluster 1, seemed concerned with pleasing the other character ($\delta = 4.15$) and seemed to act as if they had high prior entropy as shown by distribution of the lower inverse temperatures τ^{-1} in Figure 3a. Behaviourally, this can be seen from the lack of observable differences between all certainty levels in Figure 3b. We can see that the full model recovers this pattern of behaviour pretty accurately. In contrast, participants in cluster 2 seemed less concerned with pleasing the other character overall ($\delta = 1.98$), and seemed to be more sensitive to certainty of the other character as all τ^{-1} tended to be higher than for participants in cluster 1. Behaviourally, we see that when the characters were *uncertain*, their behaviour was consistent with cluster 1. However, for *certain* and *immutable* characters, we see an increase in pursuing self-interest, reaching a maximum with the *immutable* character when their trust was *low*. While the model should theoretically accommodate this pattern, we clearly see a significant decrease in r^2 for cluster 2 (see Table 1) and future work should seek to address this drop.

Discussion

As predicted, interacting with a trusting character made participants less prosocial. They cared less about making a good impression or granting requests when the other character already held a positive opinion, even if that opinion could change quickly. This highlights how presentational social goals, i.e. focused on others' beliefs about us rather than intrinsic concern for their well-being, can lead individuals to prioritize personal interests when social and personal goals conflict. Participants exhibited two distinct strategies. Cluster 1 members favored their own interests more often when they perceived the other person to be trusting, showing little concern for how their actions influenced the other's opinions. Interestingly, this behaviour seemed to interpolate between intrinsic social goals (valuing others' rewards regardless of beliefs) and presentational goals (caring about others' opinions only when they believe they can change them). In contrast, Cluster 2 members acted similarly with trusting characters but were highly sensitive to their ability to influence beliefs. They showed little interest in regaining trust when they believed their efforts would be ineffective, leading to more self-interested behavior with distrustful, unyielding agents.

Model extensions and limitations While we chose to isolate the contribution of other's beliefs to action selection by choosing a dictator game, the model can easily be extended to other economic games such as the ultimatum or trust game (Aina et al., 2020; Barnby et al., 2025; Oosterbeek, Sloof, & van de Kuilen, 2004). In those, dynamics become more complex but also more realistic in that current changes in the belief states of others can lead to direct consequences on one's future personal interests. Additionally, we currently do not let agents plan further than one step ahead, which greatly limits their flexibility. Letting agents plan should allow us to model more complex interactions which unfold over time, where short term losses can be weighed against long term gains (T. Chen, Houlihan, Chandra, Tenenbaum, & Saxe, 2024; Houlihan et al., 2023). This study has several limitations which caveat its findings. First, matching trust levels with facial expressions is imprecise: one could easily trust someone while being displeased with them. Second, our measure for certainty lacked specificity in that it was simply measured by the entropy of the prior of a neighbour. However, an unwavering belief could also stem from a close relative being certain about my intentions due to the number of years spent with each other.

Conclusion

We presented a computational model of how individuals balance material and presentational goals in social interactions. Despite no direct cost on personal gains, participants prioritized others' beliefs and the impact of their actions on those beliefs. By framing presentational goals, like maintaining trust, as minimizing the divergence between a desired and estimated belief state of the other agent, our model captures complex decision-making patterns, showing people are less likely to signal good intentions when they believe others already trust them. This approach advances understanding of social behavior and may have potential applications in studying pathologies such as borderline personality disorder, which disrupt belief updating about others (Barnby et al., 2025).

Appendix

We find expressions for the posterior distributions over β_G given actions a. Let us consider the following joint posterior:

$$p(\alpha_G, \beta_G | a) \propto p(a | \alpha_G, \beta_G) p(\alpha_G) p(\beta_G)$$

where we have used the mean field assumption for priors. Using the utility function, note that $p(a|\alpha_G, \beta_G) = \sigma(U_B(a, \alpha_G, \beta_G))$ where now the utility of B, U_B is also a function of α and β and is simplified to be the version in equation 1. Replacing in equation 1 yields:

$$\begin{aligned} p(\alpha_G, \beta_G | a) &\propto \sigma(U_B(a, \alpha_G, \beta_G)) p(\alpha_G) p(\beta_G) \\ &\propto \exp(U_B(a, \alpha_G, \beta_G)) p(\alpha_G) p(\beta_G) \end{aligned}$$

Taking the natural logarithm on both sides for clarity:

$$\begin{aligned} \log p(\alpha_G, \beta_G | a) &\propto \log \exp(U_B(a, \alpha_G, \beta_G)) + \log p(\alpha_G) + \log p(\beta_G) \\ &= \alpha_G G_B(a) + \beta_G G_R(a) + \log p(\alpha_G) + \log p(\beta_G) \end{aligned}$$

As $p(\alpha_G, \beta_G | a) = p(\alpha_G | a) p(\beta_G | a)$, from the mean field assumption, we get two separate equations for the posteriors:

$$\begin{aligned} \log p(\alpha_G | a) &\propto \alpha_G G_B(a) + \log p(\alpha_G) \\ \log p(\beta_G | a) &\propto \beta_G G_R(a) + \log p(\beta_G) \end{aligned}$$

References

Aina, C., Battigalli, P., & Gamba, A. (2020). Frustration and anger in the Ultimatum Game: An experiment. *Games and Economic Behavior*, *122*, 150–167. doi: 10.1016/j.geb.2020.04.006

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. doi: 10.1038/s41562-017-0064

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

Barnby, J. M., Nguyen, J., Griem, J., Wloszek, M., Burgess, H., Richards, L., … Fonagy, P. (2025). Self-other generalisation shapes social interaction and is disrupted in borderline personality disorder. *eLife*, *14*. (Publisher: eLife Sciences Publications Limited) doi: 10.7554/eLife.104008.1

Battigalli, P., & Dufwenberg, M. (2007). Guilt in Games. *American Economic Review*, *97*(2), 170–176. doi: 10.1257/aer.97.2.170

Battigalli, P., & Dufwenberg, M. (2022). Belief-Dependent Motivations and Psychological Game Theory. *Journal of Economic Literature*, *60*(3), 833–82. doi: 10.1257/jel.20201378

Bicchieri, C., Xiao, E., & Muldoon, R. (2011). Trustworthiness is a Social Norm, but Trusting is Not. *Politics, Philosophy and Economics*, *10*(2), 170–187. (Publisher: Sage) doi: 10.1177/1470594x10387260

Camerer, C. F., & Thaler, R. H. (1995). Anomalies: Ultimatums, Dictators and Manners. *Journal of Economic Perspectives*, *9*(2), 209–219. doi: 10.1257/jep.9.2.209

Chandra, K., Chen, T., Li, T.-M., Ragan-Kelley, J., & Tenenbaum, J. (2024). Cooperative Explanation as Rational Communication. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46). doi: 10.31234/osf.io/bmknu

Chandra, K., Li, T.-M., Tenenbaum, J., & Ragan-Kelley, J. (2023). *Acting as Inverse Inverse Planning.* (arXiv:2305.16913 [cs]) doi: 10.1145/3588432.3591510

Chandra, K., Li, T.-M., Tenenbaum, J. B., & Ragan-Kelley, J. (2024). Storytelling as Inverse Inverse Planning. *Topics in Cognitive Science*, *16*(1), 54–70. doi: 10.1111/tops.12710

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*(C), 88–97. (Publisher: Elsevier)

Chen, T., Houlihan, S. D., Chandra, K., Tenenbaum, J., & Saxe, R. (2024). Intervening on Emotions by Planning Over a Theory of Mind. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46).

Cushman, F. (2024). Computational Social Psychology. *Annual Review of Psychology*, *75*(Volume 75, 2024), 625–652. (Publisher: Annual Reviews) doi: 10.1146/annurev-psych-021323-040420

Fehr, E., & Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. (Publisher: Oxford University Press)

Guala, F., & Mittone, L. (2010). Paradigmatic experiments: The Dictator Game. *The Journal of Socio-Economics*, *39*(5), 578–584. doi: 10.1016/j.socec.2009.05.007

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

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with Theory of Mind. *Trends in Cognitive Sciences*, *26*(11), 959–971. doi: 10.1016/j.tics.2022.08.003

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. doi: 10.1098/rsta.2022.0047

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. doi: 10.1016/j.tics.2016.05.011

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(31), 8658–8663. doi: 10.1073/pnas.1601280113

Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology. General*, *152*(2), 346–362. doi: 10.1037/xge0001277

Oosterbeek, H., Sloof, R., & van de Kuilen, G. (2004). Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis. *Experimental Economics*, *7*(2), 171–188. doi: 10.1023/B:EXEC.0000026978.14316.74

Saulin, A., Horn, U., Lotze, M., Kaiser, J., & Hein, G. (2022). The neural computation of human prosocial choices in complex motivational states. *NeuroImage*, *247*, 118827. doi: 10.1016/j.neuroimage.2021.118827

Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M. S., Højsgaard, S., … Christensen, R. H. B. (2024). *afex: Analysis of Factorial Experiments.*

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

949–958). PMLR.

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

Wu, Y. E., & Hong, W. (2022). Neural basis of prosocial behavior. *Trends in Neurosciences*, *45*(10), 749–762. doi: 10.1016/j.tins.2022.06.008

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. doi: 10.1162/opmi_a_00035