A social inverse-planning model captures how relationship intimacy constrains action planning and interpretation

Alicia M. Chen (aliciach@mit.edu)[1], Ashley J. Thomas[2], Joshua B. Tenenbaum[1] & Rebecca Saxe[1]
[1]Department of Brain and Cognitive Sciences, MIT
[2]Department of Psychology, Harvard University

Abstract

Cognitive models of inverse planning typically treat a planner's utilities as functions of physical properties. However, many everyday actions are also shaped by utilities that are based on sociological properties, like social relationships. In this research, we extend inverse planning models to include social relationships as entities that change the utilities of actions. Using food sharing as a case study, we show that models require this sociological structure to capture human observers' judgments about actions, desires, and relationships. This work situates inverse planning models within everyday social interactions, where ongoing relationships constrain how people plan and interpret actions.

Keywords: inverse planning; theory of mind; social relationships; social cognition

Introduction

John and Sarah walk 16 blocks across town to get a bowl of noodles from Ramen House, passing dozens of restaurants along the way. From this simple scenario, human observers can infer that John and Sarah already *know* about Ramen House (its location, hours and menu) and *prefer* Ramen House noodles to other food options. How observers infer internal states like *knowledge* and *desire* from action is the classic problem of theory of mind reasoning (Dennett, 1989; Gopnik & Meltzoff, 1997; Wellman, 2014). A standard and effective cognitive model of this reasoning is "inverse planning" (C. L. Baker et al., 2009, 2017; Jara-Ettinger et al., 2016). Inverse planning assumes that agents choose actions rationally, given their beliefs and desires, to maximize reward and minimize costs, and uses observed actions to infer unobserved beliefs and desires. In the standard examples, rewards arise from outcomes like eating a desired food, and costs arise from properties like distance traveled. That is, the agents' internal states all directly refer to properties of the physical world.

Yet these examples leave out a fundamental class of rewards and costs that influence all social actions, including any time two people eat together. These rewards and costs arise from the sociological world of relationships and are not reducible to physical properties. Just as distance, effort, and outcomes shape which actions are feasible or worthwhile, social relationships shape which actions are comfortable, rewarding, or even conceivable (Davis et al., 2026; Jara-Ettinger & Dunham, 2024; Thomas, 2024).

For example, imagine that when they get to Ramen House, John and Sarah share a single bowl of noodles using the same chopsticks. From this simple scenario, observers can infer not only how much John and Sarah like *noodles*, but also how much they like *each other*. This is because when people plan whether and how to share food, key aspects of utilities depend on their relationship. Sharing food from the same plate, drinking from the same cup, and biting from the same part of the food, all feel comfortable only in relatively intimate relationships (Miller et al., 1998; Thomas et al., 2022). These actions are common between family members, romantic partners, or close friends, but not between students and teachers, colleagues, or casual neighbors.

These observations suggest that constraints on social actions arise from both the physical environment (e.g., distance, obstacles, effort) and the sociological environment (e.g., relationship structure). Existing inverse planning models capture how agents select actions given beliefs about physical constraints, and how observers jointly infer agents' internal states along with unobserved physical properties such as invisible barriers or food trucks that have closed. Here, we extend this framework by developing a social inverse planning model that treats sociological structure as a parallel source of constraint and inference. Incorporating sociological structure in this way enables joint inference over both physical and sociological constraints, capturing a broader range of everyday reasoning. It also explains how sociological structure systematically shifts inferences about the standard targets of inference – agents' desires – in inverse planning models.

Over three experiments, human observers make inferences about agents' expected actions, desires, and relationships. Our results show that standard inverse planning models that consider only physical costs and constraints cannot capture the full patterns of human behavior. Instead, people's inferences are closely captured by a social inverse planning model that considers how intimacy modulates the costs of sharing food.

Computational framework

We model social actions as taken by a joint "we-agent," treating the dyad as a collective actor with shared goals and mutual understanding (Chater et al., 2022; Tomasello et al., 2005). This approach contrasts with prior work that segments eating food into distinct roles (a sharer and a receiver) to model how a food receiver infers a sharer's beliefs about the relationship (Hung et al., 2022). By treating the dyad as a joint actor, our approach captures the shared background assump-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

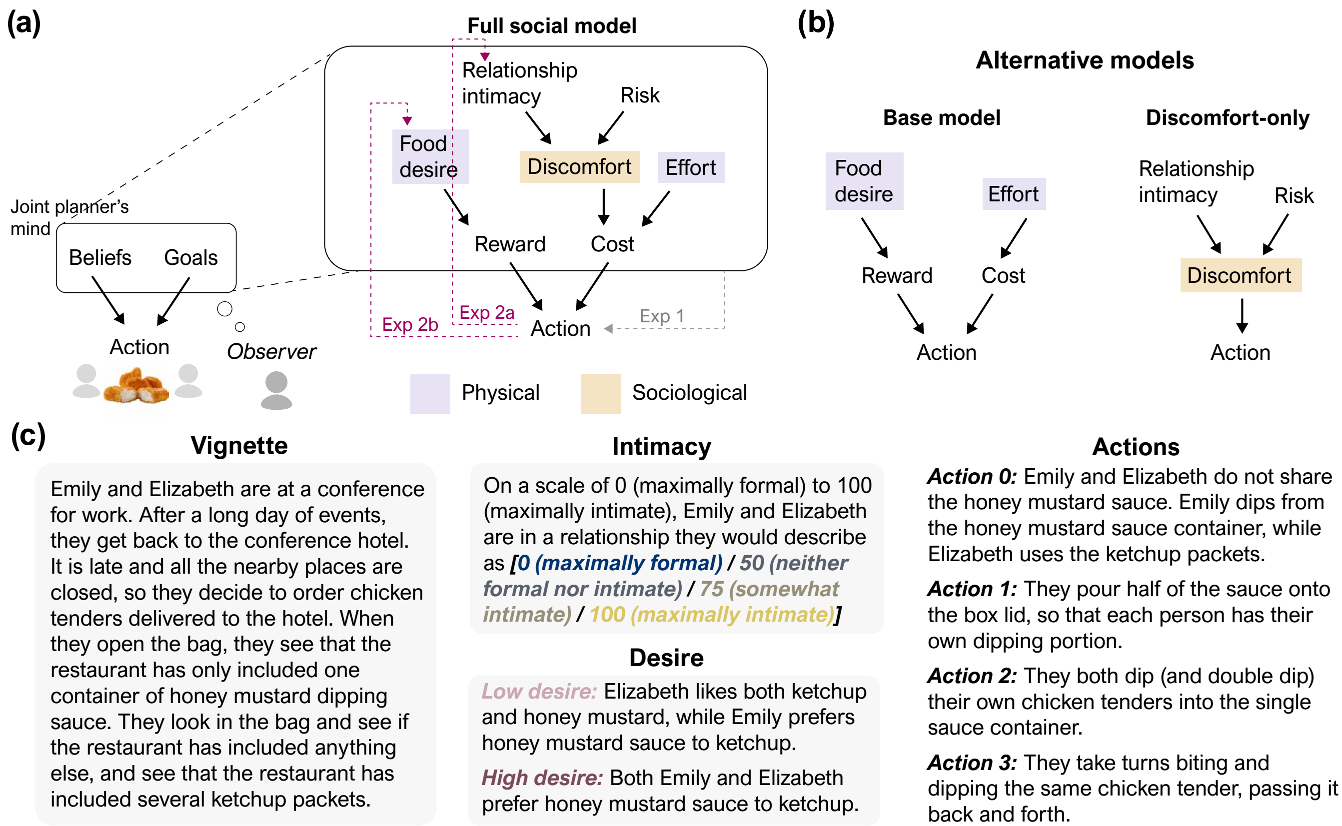

Figure 1: Overview of the social inverse planning framework and experimental design.

tions characteristic of stable dyadic relationships and abstracts away from explicit bargaining processes (Clark, 1996; Fiske, 1991). In this way we incorporate relational structures analogous to physical structure in the standard models – as entities that shape utility components and are part of common knowledge. We first describe the general framework here, and then provide experiment-specific details in later sections.

Standard inverse planning models assume that agents choose actions to maximize expected utility, defined as the reward of achieving a goal minus the costs incurred (C. L. Baker et al., 2009, 2017; Jara-Ettinger et al., 2016). The actor chooses an action proportional to its total utility, which is a weighted sum of the individual utility components. Observers can then use Bayesian inference to recover the agent's internal states from observed actions.

In standard applications, rewards are determined directly by the agent's desires, and costs arise from physical properties. For example, rewards include eating a preferred food (C. L. Baker et al., 2017; Jern & Kemp, 2014), obtaining or lifting an object (Xiang et al., 2023; Zhi-Xuan et al., 2020), or getting to a specific location (C. L. Baker et al., 2009). Costs include traveling further and across more challenging or aversive terrain (C. L. Baker et al., 2009, 2017; Jara-Ettinger et al., 2020), lifting a heavier object (Xiang et al., 2023), or jumping higher (Jara-Ettinger et al., 2020; Liu et al., 2017). In prior models, these rewards and costs are manually specified by researchers, under the assumption that such features are common knowledge.

We extend this framework by incorporating how the utilities of actions also include aspects related to social relationships (Figure 1a). As a case study, we incorporate in the models how relationship intimacy changes the *discomfort* – a kind of cost – associated with performing risky or vulnerable actions like sharing saliva. This discomfort depends both on the risk of saliva transfer and on the intimacy of the relationship.

We test the model in the domain of eating and sharing food, which is a familiar domain where the effects of physical and sociological utilities can be experimentally dissociated (we can separate *whether* the two people eat the food, from *how* they eat it). This enables direct comparison with the standard models, which have also primarily been tested in the domain of food preferences. The resulting framework can recover both desires and relationships from observed actions.

Transparency and openness

Preregistration All procedures and alternative models were preregistered[1]. The reported model comparison deviated

[1] Experiment 1: https://osf.io/9ze34/ Experiment 2a: https://osf.io/74pg8/ Experiment 2b: https://osf.io/jgvyq/

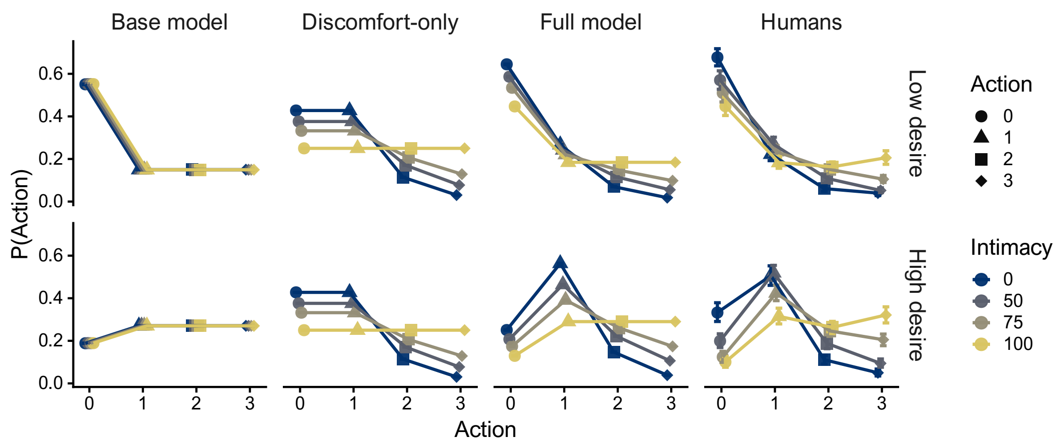

Figure 2: Experiment 1 results and best-fitting model predictions. Error bars are 95% bootstrapped confidence intervals.

slightly from the preregistrations; see the individual preregistration links for more detail.

Implementation details The experiments were created using jsPsych (de Leeuw, 2024; de Leeuw et al., 2023). The cognitive models were implemented in memo (Chandra et al., 2025). The cognitive models were implemented manually; AI tools (Claude Code and Cursor) were used to aid the generation of other code for the project. All AI generated code was validated and tested by the first author. All code and anonymized data are available at https://github.com/aliciamchen/social-inverse-planning_cogsci-26.

Experiment 1: Forward planning

In Experiment 1, we tested how observers use sociological and physical properties to predict others' actions.

Methods

Participants For all experiments, participants were recruited on Prolific and pre-screened to be adult fluent English speakers from the United States. Participants were excluded if they failed an attention check and two memory checks. For Experiment 1, we recruited 118 (70 female, 46 male, 1 non-binary; ages 18–77, M(SD) age = 42.3(13.7)) participants, and 3 participants were excluded. Participants were paid $6.25 for completing the experiment, which took an estimated 25 minutes. Participants completing one experiment were excluded from subsequent experiments. Participants gave informed consent, and all procedures were approved by the MIT Committee for the Use of Humans as Experimental Subjects.

Materials We wrote 16 scenarios describing situations in US cultural contexts where two people might plausibly share food (e.g., cake at a birthday party, a hot dog at a basketball game, drinks at a house party; see Figure 1c for an example; all scenarios available in the project repository). The gender pairing of the characters' names varied across scenarios.

Procedure Participants read vignettes describing food-sharing scenarios and rated the probability that two characters will take each of four possible actions, ranging from not sharing the food at all to sharing in ways that involve increasing risk of saliva transfer.

We manipulated two factors: how much the two characters want the food, and the intimacy of their relationship (Figure 1c). In the 'high desire' condition, both characters want to eat the food (e.g., both like the food, both are hungry). In the 'low desire' condition, one character wants the food less (e.g., prefers another food, is not hungry). Relationship intimacy was described on a 0–100 scale, ranging from maximally formal (0) to maximally intimate (100), and sampled at four levels (0, 50, 75, 100). The resulting design was thus 2 (desire: low vs. high) x 4 (intimacy: 0, 50, 75, 100). Each participant saw all 16 scenarios; the assignment of scenario to condition was balanced within and across participants.

On each trial, participants allocated probability mass across four actions. Action 0 describes the two characters not sharing the food (e.g., one person eats the food while the other eats something else). Action 1 describes them sharing the food in a way that does not risk saliva transfer (e.g., pouring separate portions). Action 2 and 3 are ways to share the food with increasing risk of saliva transfer (see Figure 1c).

Model specification

In the full social model (Figure 1a), a joint actor chooses an action $a \in \{a_0, a_1, a_2, a_3\}$ proportional to its total utility, following a softmax choice rule with inverse temperature α. The total utility of an action a given motivational state (desire) $s \in \{\text{low}, \text{high}\}$ and intimacy $I \in [0, 1]$ is:

$$U(a|s, I) = w_v \cdot V(a|s) - w_d \cdot d(a|I) - w_e \cdot e(a) \quad (1)$$

To generate model predictions, we assigned each action values based on the experimental design, reflecting qualitative distinctions that are explicitly encoded in the scenarios. We describe each component of the utility below.

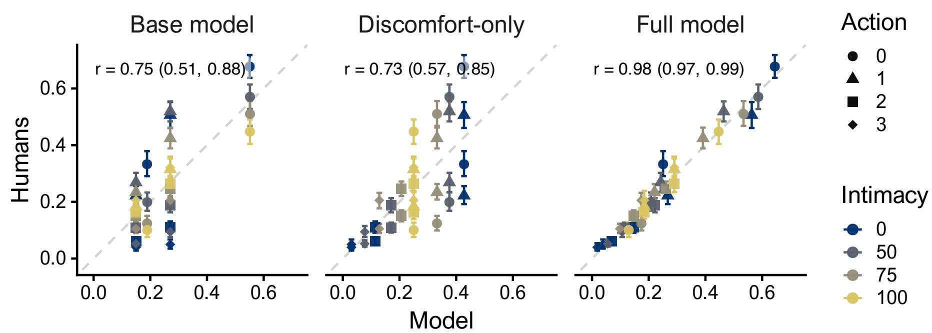

Figure 3: Model-human correlations in Experiment 1. Error bars are 95% bootstrapped confidence intervals.

Reward The reward $V(a|s)$ of an action depends on the action and the desire state. In the 'high desire' state, sharing actions (a_1–a_3) serve the goal of jointly eating the food: $V(a|s = \text{high}) = [0, 1, 1, 1]$ for actions a_0 through a_3. In the 'low desire' state, the same sharing actions have negative utility: $V(a|s = \text{low}) = [0, -1, -1, -1]$.

Discomfort The *discomfort* of taking an action depends on both the risk of transferring saliva via the action and the intimacy of the relationship. We set a saliva-transfer risk $\rho = [0, 0, 1, 2]$ for a_0 through a_3. Intimacy reduces the discomfort caused by risk, and the shape is controlled by γ:

$$d(a|I) = \rho(a) \cdot (1 - I)^{\gamma} \tag{2}$$

Effort Finally, the effort $e = [0, 1, 1, 1]$ captures the baseline physical cost of sharing food versus not sharing.

Alternative models

We considered two alternative models (Figure 1b), each inspired by part of the prior literature:

- A 'Base' model that incorporates only the reward for eating the food and the effort of sharing.

$$U_{\text{base}}(a, s) = w_v \cdot V(a|s) - w_e \cdot e(a) \tag{3}$$

- A 'Discomfort-only' model that captures only how intimacy mitigates discomfort from risky actions (Miller et al., 1998).

$$U_{\text{discomfort-only}}(a|I) = -w_d \cdot d(a|I) \tag{4}$$

Results

Table 1: Experiment 1 model comparison. ΔAIC is computed relative to the Full model (positive = worse than Full).

Model	AIC	ΔAIC	r (95% CI)
Full	4349.11	0	0.98 (0.97, 0.99)
Discomfort-only	4624.28	275.17	0.73 (0.57, 0.85)
Base	4707.94	358.83	0.75 (0.51, 0.88)

We fit each model to participants' probability judgments by maximum likelihood using stochastic gradient descent, fixing $\alpha = 1$ for identifiability. We generated predictions using leave-one-scenario-out cross-validation, refitting on 15 scenarios and predicting the held-out scenario. We compared models by in-sample AIC and out-of-sample Pearson correlation between predictions and human cell means, pooled across CV folds.

The experimental design allows us to test two predictions: whether food desire reverses preferences between not sharing (Action 0) and no-risk sharing (Action 1), and whether relationship intimacy selectively increases tolerance for saliva-transfer risk, raising the likelihood of higher-risk sharing (Actions 2 and 3).

The Base model captures the desire-based reversal between Actions 0 and 1 but is insensitive to intimacy. The Discomfort-only model captures how intimacy increases risk tolerance, but cannot represent desire and therefore does not distinguish between sharing and not sharing. Only the full social model (best-fitting params $w_v = 0.85$, $w_d = 1.34$, $w_e = 0.03$, $\gamma = 0.86$) captures both patterns (Figure 2, Figure 3). These results show that representing relational structure is necessary to capture forward planning judgments in this domain.

Experiment 2: Inverse planning

In Experiment 2, we tested whether the fits from Experiment 1 generalize to inverse planning. Two complementary experiments test both inferential targets, intimacy (Experiment 2a) and food desire (Experiment 2b).

Methods

Participants For Experiment 2a, we recruited 117 participants (56 female, 56 male, 3 non-binary, 2 abstain to respond; ages 19–78, M(SD) age = 44.3(13.9)) and excluded 3 participants. For Experiment 2b, we recruited 120 participants (62 female, 51 male, 5 non-binary; ages 19–74, M(SD) age = 41.1(13.2)) and did not exclude any participants. Participants were paid \$5 for completing the experiment, which took an estimated 20 minutes to complete.

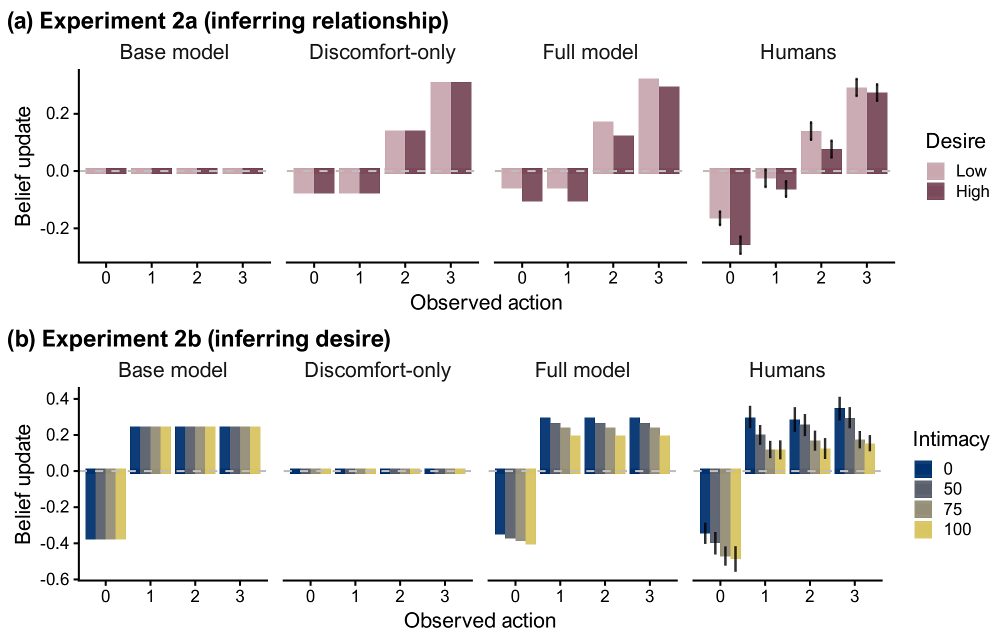

Figure 4: Experiment 2 results and best-fitting model predictions. Error bars are 95% bootstrapped confidence intervals.

Procedure The scenarios were the same as in Experiment 1 (Figure 1c), but the experiments followed the inverse structure. On each trial, participants provided ratings both before and after observing the characters take a specific action; the primary dependent variable is the belief update (posterior minus prior). In Experiment 2a, participants rated relationship intimacy (0-100 scale) given desire (low vs. high) and observed action (0-3), yielding a 2×4 design. In Experiment 2b, participants rated the relative probability of the low-desire versus high-desire state, given intimacy (0, 50, 75, or 100) and observed action (0-3), yielding a 4×4 design.

Model specification

Upon observing an action, an observer who knows one latent variable (desire state or intimacy) can infer the other by inverting the forward model:

$$P(I|a,s) \propto P(a|s,I) \cdot P(I|s) \tag{5}$$
$$P(s|a,I) \propto P(a|s,I) \cdot P(s|I), \tag{6}$$

where $P(a|s,I)$ is the forward planning likelihood from Experiment 1. Because the dependent variable is belief update, we set uniform priors so that predicted belief update directly reflects the diagnostic value of the observed action.

For each model, we froze the best-fitting actor parameters obtained from Experiment 1 to generate the inverse predictions, and fit only an additional parameter, α_{obs} (so that the effective likelihood term is $P(a|s,I)^{\alpha_{\text{obs}}}$), by minimizing the mean squared error between the model's predicted belief update and the human belief update. This parameter captures how sharply people weigh higher-likelihood states. The reported fits and predictions are out of sample, following the cross-validation procedure of Experiment 1.

Results

Table 2: Experiment 2 model comparison: out-of-sample per-trial ΔMSE relative to the Full model, pooled across folds.

Model	ΔMSE (Exp 2a)	ΔMSE (Exp 2b)
Full	0	0
Discomfort-only	0.001	0.080
Base	0.030	0.003

Across both experiments, observers' belief updates reflected joint reasoning about reward, cost, and relationship context, in ways that the alternative models could not capture (Figure 4). The Base model captures only the basic inference of how sharing (actions 1–3) increases inferred desire and non-sharing (action 0) decreases it. Conversely, the Discomfort-only model predicts greater intimacy from higher-risk sharing (actions 2–3), but cannot account for desire or how it interacts with intimacy.

By contrast, the social model captures systematic 'explaining-away' effects between desire and intimacy. Risky sharing is informative about intimacy primarily when desire is low; when desire is high, the same behavior is less diagnostic because motivation to eat already provides an alternative explanation for accepting discomfort. On the other hand, choosing not to share despite really wanting the food (high desire) is strongly diagnostic of a more formal relationship (Figure 4a). A parallel pattern emerges for desire inference. When intimacy is low, risky sharing strongly increases in-

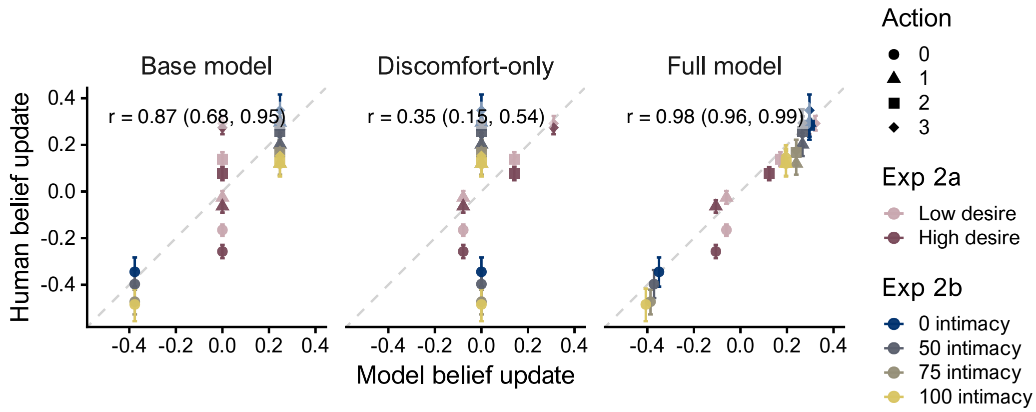

Figure 5: Model-human correlations for Experiment 2. Error bars are 95% bootstrapped confidence intervals.

ferred desire; when intimacy is high, the same risky actions are less informative because the relationship already explains why someone would be comfortable sharing (Figure 4b).

Across both Experiment 2a and Experiment 2b, the full social model achieved the lowest MSE, though the margin over the next-best alternative was small. The full model is, however, the only model that qualitatively captures the explaining-away pattern across both inferential targets (Figure 5).

Discussion

The present work extends inverse planning models by treating sociological structure – specifically, relationship intimacy – as a factor that systematically shapes the utilities of action. Much of everyday social reasoning involves observing people who are acting together, whether on a date, at a dinner party, or in a meeting. In these contexts, the utilities of the actions depend not only on physical properties but also on the relationship between the actors.

Prior work has incorporated relationships into models of social reasoning in several ways, including by representing how agents weigh others' utilities (Jern & Kemp, 2014; Kleiman-Weiner et al., 2017; Powell, 2022; Ullman et al., 2009), their attitudes toward others (Wang & Jara-Ettinger, 2025), or by specifying behaviors consistent with particular roles or relationships (A. Baker et al., 2025; Davis et al., 2026). These approaches capture important aspects of social reasoning, but typically treat relationships as separate objects of reasoning rather than structures that directly constrain existing utilities. By contrast, our approach allows relational structure to interact with – and be compared directly to – the physical variables that have traditionally constituted inverse planning models.

Modeling two characters as a joint "we-agent" reflects the shared assumptions that support coordination in stable relationships. This approach is well suited to modeling how third-party observers reason about interacting dyads in contexts where relational structure can be treated as stable and mutually known. At the same time, it does not capture intra-relationship dynamics such as negotiating roles or managing uncertainty about the relationship (Chuey & Gweon, 2025; Hung et al., 2022; Wu et al., 2021; Xiang et al., 2023). These processes likely play an important role in contexts where relationships are uncertain or evolving.

Although we studied food sharing as a case study, the same framework should generalize to other social actions where relationship structure shapes what actions feel appropriate, comfortable, or desirable. These include asking for help, self-disclosure, gift giving and generosity, and teasing (Chen & Saxe, 2025; Fiske, 1991; Keltner et al., 1998; Nadler et al., 1985; Taylor & Altman, 1975). Intimacy is also only one dimension of social structure: group membership, status and hierarchy, or culturally specific roles may similarly influence how agents trade off rewards and costs (Fiske, 1991; Jara-Ettinger & Dunham, 2024; Thomas, 2024). These factors can be incorporated within the same framework by specifying how they alter utility components.

More broadly, these results suggest that inverse planning models should explicitly represent how social context shapes utilities. When models succeed without relational variables – even in minimal grid-world paradigms – this may reflect the implicit incorporation of relational assumptions into other utility terms, or the selective omission of it when it is not relevant for the current judgment (Wong et al., 2025). We argue that making these assumptions explicit is necessary for understanding both sociological structure and the commitments of existing models.

Acknowledgments

We thank Charlston Stovall for assistance with the stimuli, and Marlene Berke and Fernanda de la Torre for helpful discussions. This work was funded by the Simons Foundation Autism Research Initiative.

References

Baker, A., Sharma, K., Dunham, Y., & Jara-Ettinger, J. (2025). People use mixed strategies to make efficient but structured inferences about agents in roles. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

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.

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). *Proceedings of the ACM on Programming Languages*, *9*(OOPSLA2), 784–814.

Chater, N., Zeitoun, H., & Melkonyan, T. (2022). The paradox of social interaction: Shared intentionality, we-reasoning, and virtual bargaining. *Psychological Review*, *129*(3), 415.

Chen, A. M., & Saxe, R. (2025). Expectations of reciprocal generosity are specific to equal relationships.

Chuey, A., & Gweon, H. (2025). Theory of minds: Early understanding of interacting minds. *Annual Review of Developmental Psychology*, *7*(1), 91–115.

Clark, H. H. (1996). *Using language*. Cambridge university press.

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*.

de Leeuw, J. R. (2024). Datapipe: Born-open data collection for online experiments. *Behavior Research Methods*, *56*(3), 2499–2506.

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.

Dennett, D. (1989). *The intentional stance*. MIT press.

Fiske, A. P. (1991). *Structures of social life: The four elementary forms of human relations: Communal sharing, authority ranking, equality matching, market pricing.* Free Press.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories.* The MIT Press.

Hung, M. S., Thomas, A. J., Radkani, S., Tenenbaum, J., & Saxe, R. (2022). Modeling risky food sharing as rational communication about relationships. *Proceedings of the annual meeting of the cognitive science society*, *44*(44).

Jara-Ettinger, J., & Dunham, Y. (2024). The institutional stance.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naive utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, *20*(8), 589–604.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334.

Jern, A., & Kemp, C. (2014). Reasoning about social choices and social relationships. *Proceedings of the annual meeting of the cognitive science society*, *36*(36).

Keltner, D., Young, R. C., Heerey, E. A., Oemig, C., & Monarch, N. D. (1998). Teasing in hierarchical and intimate relations. *Journal of personality and social psychology*, *75*(5), 1231.

Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition*, *167*, 107–123.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Miller, L., Rozin, P., & Fiske, A. P. (1998). Food sharing and feeding another person suggest intimacy; two studies of american college students. *European Journal of Social Psychology*, *28*(3), 423–436.

Nadler, A., Peri, N., & Chemerinski, A. (1985). Effects of opportunity to reciprocate and self-esteem on help-seeking behavior. *Journal of Personality*, *53*(1), 23–35.

Powell, L. J. (2022). Adopted utility calculus: Origins of a concept of social affiliation. *Perspectives on Psychological Science*, *17*(5), 1215–1233.

Taylor, D. A., & Altman, I. (1975). Self-disclosure as a function of reward-cost outcomes. *Sociometry*, 18–31.

Thomas, A. J. (2024). Cognitive representations of social relationships and their developmental origins. *Behavioral and Brain Sciences*, 1–53.

Thomas, A. J., Woo, B., Nettle, D., Spelke, E., & Saxe, R. (2022). Early concepts of intimacy: Young humans use saliva sharing to infer close relationships. *Science*, *375*(6578), 311–315.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, *28*(5), 675–691.

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*.

Wang, Z., & Jara-Ettinger, J. (2025). When seating matters: Modeling graded social attitudes as bayesian inference. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University Press.

Wong, L., Collins, K. M., Ying, L., Zhang, C. E., Weller, A., Gerstenberg, T., O'Donnell, T., Lew, A. K., Andreas, J. D., Tenenbaum, J. B., et al. (2025). Modeling open-world cognition as on-demand synthesis of probabilistic models. *arXiv preprint arXiv:2507.12547*.

Wu, S. A., Wang, R. E., Evans, J. A., Tenenbaum, J. B., Parkes, D. C., & Kleiman-Weiner, M. (2021). Too many

cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, *13*(2), 414–432.
Xiang, Y., Vélez, N., & Gershman, S. J. (2023). Collaborative decision making is grounded in representations of other people's competence and effort. *Journal of Experimental Psychology: General*, *152*(6), 1565.
Zhi-Xuan, T., Mann, J., Silver, T., Tenenbaum, J., & Mansinghka, V. (2020). Online bayesian goal inference for boundedly rational planning agents. *Advances in neural information processing systems*, *33*, 19238–19250.