People balance joint reward, fairness and complexity to develop social norms in a two-player game

Dhara Yu, Bill D. Thompson
University of California, Berkeley
{dharakyu, wdt}@berkeley.edu

Abstract

Social norms are a hallmark of human social intelligence, yet the reasoning processes involved in norm formation have been difficult to capture with traditional modeling frameworks. We developed a computational model of norm formation as joint planning via theory-of-mind. The model is designed to capture the distinctively human ability to flexibly develop more complex norms in more complex situations, via simulation of joint decision-making with other agents over an extended time horizon. We evaluated the predictions of the model against participant interactions in a 2-player iterated decision-making task. Across 3 conditions our model captured the way participants balanced joint reward, fairness, and complexity when forming norms.

Keywords: social norms; theory of mind; joint planning

Introduction

Human cooperation relies on the ability to develop and reason about social norms (Tomasello, 2019). Norms are group-wide expectations about how individuals in a group should act in a particular context, such that individuals are expected to both conform to a norm and to enforce norm compliance among others (Bicchieri, 2005).

One function of norms is to help people solve simple coordination problems, such as choosing which side of the road to drive on. But the power of norms goes beyond simple choices, offering solutions to complex situations that extend over time and involve conflicting incentives for individuals and groups (Hawkins, Goodman, & Goldstone, 2019). For example, consider two housemates who each own a car but share a single parking spot. This is a more complex coordination problem, because what is optimal for one person—parking in the shared spot every single time—is bad for the other. Yet the housemates could quickly and intuitively develop a better solution, such as a norm of alternating weeks when each person can use the garage.

The flexibility with which people form structured norms is difficult to capture with existing modeling frameworks. For example, classical game theoretic models of interdependent choice based on utility maximization struggle to account for norms of alternation (Helbing, Schönhof, Stark, & Hołyst, 2005). Evolutionary game theory models have shown that agents pre-programmed to play with the contingently cooperative "tit-for-tat" strategy can outcompete other agents (Axelrod, 1984), but such models build strategies into the model directly, offering limited insight into the cognitive and interpersonal mechanisms through which complex strategic norms arise (Gavrilets, Tverskoi, & Sánchez, 2024). The adaptive nature of norm formation suggests that norms are rooted in general cognitive principles, and in particular, inferential social reasoning about what is good for the agents involved, what is fair, as well as the *simplicity* of a solution: an agreement to swap parking spots every 3 days for one month, and then swap every 11 days for three months is intuitively less appealing than alternating weekly.

One way to capture these key cognitive principles is to view normative reasoning as an extension of theory of mind - the ability to make inferences about the mental states of others. This theoretical framework is useful because it describes how people make predictions about how others will behave in future interactions, an essential component of norm formation. Within this perspective, the capacity to make good predictions is rooted in a person's ability to simulate joint decision-making with other agents, which in turn requires reasoning about the latent beliefs and desires that give rise to action (Ho, Saxe, & Cushman, 2022). By reasoning about the mental states of other interacting agents, people can approximate the sequence of decisions most likely to be conceived by others as mutually beneficial (Misyak & Chater, 2014; Levine, Chater, Tenenbaum, & Cushman, 2023), enabling convergence to a systematic pattern of behavior and the emergence of a norm.

We developed a formal model of normative reasoning as joint planning via theory-of-mind inference. Our model combines classical formalisms in planning and decision-making with cognitive models of theory of mind and joint intentionality (Kleiman-Weiner, Ho, Austerweil, Littman, & Tenenbaum, 2016), and integrates a notion of action sequence complexity as a regularizer over the combinatorial space of joint plans. The model captures the hypothesis that people trade off the joint reward, fairness and simplicity of a candidate joint plan to generate a strategic plan over an extended time horizon in iterated social decision-making settings. Our model predicts that people prefer simpler norms, but can flexibly develop more complex strategies when necessary to prevent unfair or suboptimal allocations of reward.

To test the predictions of this theory, we conducted a behavioral experiment in which participants performed an iterated cooperative decision-making task in pairs. The task involved making simultaneous decisions with a partner about

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

who chooses which option from an array of differentially rewarding choices (parking spots with different prices in a virtual parking lot). To perform well at this task without communicating, participants needed to make inferences about the intentions and desires of their partners to develop a norm, in the form of a shared, systematic strategy for making choices. We manipulated the reward structure to induce conflicts between joint reward, fairness and complexity, enabling us to study the contingencies that influence how people develop norms. We evaluated the predictions of our model against the behavioral data, finding that our theory-of-mind model better predicted the distribution of participant norms than did simpler models that lack mechanisms to reason about fairness or complexity.

Computational framework

In this section we formalize how people generate a probability distribution over joint plans of action by trading off the reward associated with a joint plan against its complexity. We define our problem setting using the stochastic game formulation, a generalization of a Markov decision process (Littman, 1994). A 2-player stochastic game is defined as $\{S, A_1, A_2, U_1, U_2, T, \gamma\}$, where S is the joint state space for the 2 agents, and $A_1 \times A_2$ is the joint action space. $U_i(s, a_1, a_2)$ for $s \in S, a_1 \in A_1, a_2 \in A_2$ represents the reward earned for agent i in state s with agent 1 taking action a_1 and agent 2 taking action a_2. $T(s'|s, a_1, a_2)$ represents the probability of entering state s' from state s, with agents taking actions a_1, a_2. γ represents the discount factor.

This formulation can also account for *iterated* decisions. Past work has typically modeled how agents determine optimal policies within a single interaction, making it difficult to capture strategies realized over multiple interactions. To address this we formulated the action space as one of decisions over multiple time steps, i.e. over classes of joint plans, affording the flexibility to model more complex strategies. A joint plan τ over t interactions is represented as a state in S: $\tau = [(a_1^1, a_2^1), ..., (a_1^t, a_2^t)]$, where (a_1^k, a_2^k) represents the joint action taken by agents 1 and 2 on the kth interaction.

We define the probability of a joint plan $P(\tau)$ as follows:

$$U(\tau) = \underbrace{w_j \cdot R_j(\tau) + w_f \cdot R_f(\tau)}_{\text{reward}} - \underbrace{w_c \cdot C(\tau)}_{\text{complexity}}$$

$$P(\tau) \propto \exp(\beta \cdot U(\tau))$$

The utility function U is comprised of two components, a reward term and a complexity term. These two terms capture how individuals, for a candidate joint plan, weight the extent to which the plan results in optimal allocation of reward—for the overall group and between individuals—and the extent to which it is simple and cognitively efficient.

Within the reward term, $R_j(\tau)$ represents the *joint optimality*: the joint reward for all agents should they execute the given joint plan. This follows past work on modeling cooperative planning in individuals as simulating the actions of a group "we-agent" (Kleiman-Weiner et al., 2016; S. A. Wu et al., 2021). The we-agent plans over the shared state and action space of all interacting agents, representing an individual's capacity to plan with *shared intentionality* (Tomasello, Carpenter, Call, Behne, & Moll, 2005). We assume that both agents' utilities are equally weighted in computing the joint reward: $R_j(\tau) = 0.5 \cdot R_{j,1}(\tau) + 0.5 \cdot R_{j,2}(\tau)$. To compute $R_{j,i}(\tau)$, that is, the joint optimality of plan τ for agent i, we use policy iteration, which computes the optimal value function that can then be used to assign a reward to a plan.

$R_f(\tau)$ represents the *fairness* of the given joint plan: the difference in the individual rewards between agents if that plan were to be executed. Note that this is distinct from joint optimality, because a plan that is jointly optimal for multiple agents (i.e. maximizes the sum of rewards) may nonetheless result in a gap between the reward earned by each agent.

Planning over an extended time horizon surfaces a combinatorial space of candidate plans that result in equal or near-equal utility. We introduce a complexity penalty $C(\tau)$ to capture the intuition that people prefer simpler joint plans. A complexity penalty is motivated by at least two non-mutually exclusive interpretations: it can be thought of as the cognitive cost required to conceive a particular joint plan, and/or as the difficulty of coordinating the joint plan with another player. We quantified complexity using a program induction model that constructs a program that generates an observed sequence of joint actions, represented in terms of compositional operators. This model is formulated as Bayesian inference over a probabilistic context-free grammar (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Piantadosi, Tenenbaum, & Goodman, 2012), where the unit of primitive is a joint action taken by two agents. Following past work (Kleiman-Weiner et al., 2020), our model includes two compositional operators: `concat(a, b)` and `repeat(a, n)`. `concat(a, b)` combines joint actions a and b, and `repeat(a, n)` replicates the joint action a n times. As concrete examples, the sequence of joint actions $[a, a, a, a]$ is most concisely represented as the program `repeat(a, 4)`, and the sequence $[a, b, a, b]$ is most concisely represented as `repeat(concat(a, b), 2)`. The complexity penalty of a joint plan is proportional to the length, in number of operations, of the simplest program π constructed to generate it: $C(\tau) \propto |\pi(\tau)|$.

Experimental Methods

Task overview

We developed a 2-player iterated cooperative decision-making task (Figure 1). This task builds on a large literature on mixed-incentive games (Thielmann, Böhm, Ott, & Hilbig, 2021; Le Pargneux, Chater, & Zeitoun, 2023) and is designed with intuitive reward contingencies to elicit meaningful reasoning about other participants' beliefs and intentions.

Participants are informed that they must select a parking spot in a virtual parking lot over the course of several days. Different spots cost different amounts of a virtual currency (*Monetary Units*; price remained fixed over days). Participants were incentivized to minimize cost paid.[1] There were

[1]To follow the notation of the model, we construe the price of

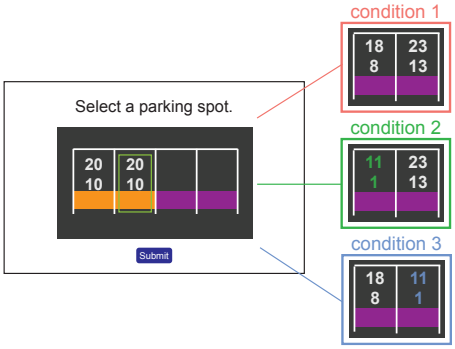

Figure 1: Task interface, showing the different reward assignments across the 3 conditions (the cost of the orange spots remains constant across the conditions). In each spot, the top number indicates the regular price and the bottom number indicates the discount price.

two zones in the lot: an orange zone and a purple zone, and two parking spots per zone. Participants were informed that if they both select a spot in the same color zone, they would receive a "group discount" and pay the discount price (visible during decision-making). Participants selected a parking spot without seeing the other person's selection, and knew that collisions incur a high price. After making their decisions, participants were shown the actions that each other took and the price each participant paid.

Experimental manipulation

We designed 3 different price configurations of the parking lot to examine whether emergent strategies reflected tradeoffs between reward, fairness, and complexity as predicted by our model (Figure 1). [2]

Condition 1: No Conflict In Condition 1, the orange spots were equal in price, while the purple spots were unequal prices. The cost of the orange spots was lower than the mean cost of the purple spots. This condition facilitates a clear strategic equilibrium: one player picking the left orange and the other player picking the right orange (we will refer to this strategy as *stable selection on orange*). This strategy maximizes joint reward and fairness, and minimizes the coordination cost (picking the same spot every time is the simplest possible process). This condition serves as a control to establish that people were capable of developing cooperative, systematic norms when a simple optimal solution is available.

Condition 2: Unavoidable Compromise Condition 2 maintains the same cost structure as in Condition 1, with one difference: the first purple spot had a regular price of 11 and a discount price of 1. This manipulation introduces a conflict between the utility terms in our model: no one strategy is optimal with respect to all terms, because the the purple spots compromise fairness and the orange spots sacrifice reward. Under these conditions, our model predicts a more diffuse set of strategies compared to Condition 1: pairs may develop a norm of *stable selection on orange* (maximally fair, minimally complex, but not jointly optimal), *stable selection on purple* (jointly optimal, minimally complex, but not fair), or *alternating selection* of the cheaper and costlier spots on purple (jointly optimal, maximally fair, but more complex). In contrast, a model without a complexity penalty would be unable to account for people's preferences for systematic strategies, and a model without a reward objective would be unable to capture the preference for higher-reward strategies.

Condition 3: Fairness vs. Complexity In Condition 3, the second purple spot has a regular price of 11 and a discount price of 1. This condition is similar to Condition 2 but with one key difference: the price gap between the two purple spots is relatively smaller. Therefore, our model predicts that people are more likely in this condition, compared to Conditions 1 and 2, to develop a norm of stable assignments on purple, because the joint reward associated with the stable purple norm is greater (compared to Conditions 1 and 2) and doing so would result in a more fair allocation of reward (compared to Condition 2). In contrast, a model without a complexity penalty would predict higher rates of complex norms or a failure to form any norm at all.

Participants and procedure

We pre-registered a target sample size of 300 participants (50 pairs per condition). Participants were recruited over multiple sessions using an algorithm that had a budget for re-recruitment to replace participants that did not complete the task (e.g. due to technical errors or waiting time limits). We excluded from analysis participants who failed to select a parking spot in the allotted time and did not finish the game, as well as participants who wrote fewer than 10 characters in a pre-game writing task. After exclusions there were 102 players (51 games) in Condition 1, 102 players (51 games) in Condition 2, and 84 players (42 games) in Condition 3.

Participants were assigned at random to one of the treatments via a block random assignment algorithm. Participants viewed instructions and had to pass a short comprehension test to advance. They were shown the parking lot of their assigned treatment and were asked to write a strategic plan describing how they would ideally play the game. After writing, participants progressed to a treatment-specific waiting room and were paired with the first available partner. They played 12 trials of the game; after each trial, participants were shown their partner's move and cost paid on the previous trial, and needed to indicate the cost they themselves paid as an attention check. The task took a median time of 12 minutes. Participants were paid a base rate of $12.50/hr; they were incentivized to minimize their overall cost in the game through a performance-based compensation bonus. Participants were informed they would play multiple trials but not told precisely how many, to induce uncertainty in the time horizon.

parking spots in terms of reward; picking the lowest-priced spot is equivalent to picking the highest reward option.

[2] Study designs, exclusion criteria and analyses were pre-registered at `https://osf.io/39fsd`.

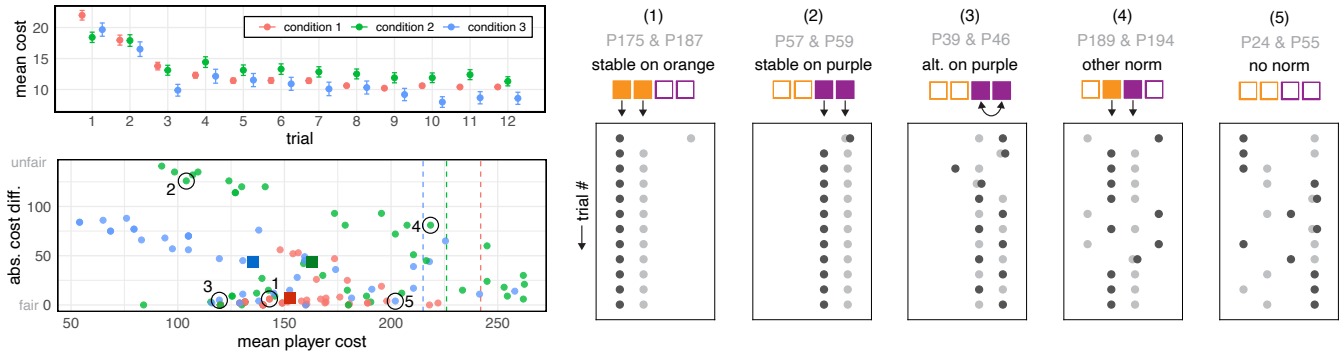

Figure 2: Left, top: mean cost paid per trial, across conditions. Error bars show standard error. Left, bottom: mean overall game cost plotted against the cost difference, for each pair. Darker squares represent the mean for all pairs in the condition. Dashed lines represented the expected mean cost for a random selection strategy. Right: example game traces, corresponding to the numbered games on the scatter plot. Each light or dark gray dot represents a participant's choice on a given trial.

Results

Behavioral results

Participants established norms Participants developed norms over the course of their interactions across all 3 conditions. The task is designed in such a way that better performance, in the form of a lower cost incurred, requires a coherent strategy on the part of the two players; thus, the cost paid by participants in a pair is a key behavioral indicator of effective norm formation. Figure 2 (left, top) shows the cost paid per trial. A Bayesian mixed-effect linear regression analysis revealed an effect of trial number on the individual cost paid, including a random effect for the group. Trial number predicted price paid ($\beta = -0.81, 95\%$ credible interval (CrI) $= [-0.92, -0.69]$); there was also an interaction effect between the trial number and the treatment in Condition 2 ($\beta = 0.28, \text{CrI} = [0.11, 0.43]$), indicating reduced decreases in cost over trials in Condition 2. These results show that participants paid less over the course of a game, indicating convergence to systematic norms.

Different reward contingencies led to different behaviors Participants' first decisions differed across conditions. Participants in Conditions 2 and 3 were more likely to select a purple spot initially, compared to Condition 1 (Condition 2: $\beta = 0.19, \text{CrI} = [0.06, 0.32]$; Condition 3: $\beta = 0.51, \text{CrI} = [0.38, 0.64]$). Over the course of the game, participants in Conditions 2 and 3 crashed more frequently compared to Condition 1 (Condition 2: $\beta = 1.00, \text{CrI} = [0.29, 1.70]$; Condition 3: $\beta = 1.25, \text{CrI} = [0.54, 2.01]$). Overall outcomes also differed between conditions (Figure 2, left, bottom). Compared to Condition 1, there was no evidence of a significant difference in the aggregate mean costs paid by participants in Condition 2 ($\beta = 10.58, \text{CrI} = [-6.73, 28.56]$) and Condition 3 ($\beta = -16.82, \text{CrI} = [-35.55, 2.06]$), even though it is in theory possible to earn a lower mean reward in those 2 conditions compared to Condition 1. However, pairs in Condition 2 and Condition 3 paid more unequal cost distributions relative to Condition 1 (Condition 2: $\beta = 36.45, \text{CrI} = [23.59, 49.64]$; Condition 3: $\beta = 35.58, \text{CrI} = [21.82, 49.56]$).

Figure 2 (right) shows examples of the types of norms that players developed over the course of a game. The five categories illustrated are 1) stable selection on orange, 2) stable selection on purple, 3) alternating selection on purple, 4) some other systematic norm, i.e. a stable or alternating on a color combination not encompassed by the first 3 categories, and 5) no apparent norm.

Strategy classification

Having established that people successfully developed norms, we analyzed the norms that formed and examined how they differed between conditions. We defined a norm as present if participant decisions were consistent with that norm for at least 2 consecutive interactions. To identify types of norms and their frequency of appearance, we developed a simple algorithm (Algorithm 1) for classifying the norms developed within pairs over the course of a game. The algorithm detects the longest consecutive sequence of pair decisions consistent with a particular strategy, and computes the proportion of the game during which the pair exhibited each norm type. Motivated by our results showing that participants converged on coherent strategies toward the end of the experiment (see Figure 2, left, top), we classified norms based on pair interactions for the final 4 trials of each game, although the conclusions are the same if analyzing the whole game.

Figure 3 shows how the types of norms that pairs developed differed across conditions. In Condition 1 (No Conflict), participants overwhelmingly converged on a norm of *stable selection on orange*, consistent with our hypothesis that participants would conceptualize that as the best strategy.

In Condition 2 (Unavoidable Compromise), the norms were more varied; compared to the Condition 1 control group, participants were less likely to exhibit stable selection on orange ($\beta = -0.46, \text{CrI} = [-0,61, -0.31]$), and more likely to fail to develop any norm ($\beta = 0.24, \text{CrI} = [0.11, 0.37]$). There was no evidence for significant differences in the frequencies of the *stable selection on purple* ($\beta = 0.13, \text{CrI} =$

Algorithm 1 Strategy classification

Input: pair decisions p

```
n ← length of p
m ← strategy for n trials; init. to null for each element
for i in [n, n−1, ..., 2] do
    S ← all subsequences of p of length i
    for each s in S do
        if there is a marked move in s then
            continue
        end if
        t ← type of strategy in s
        if t ≠ null then
            mark in m that subseq. s is strategy type t
        end if
    end for
end for
return m
```

$[-0.02, 0.28]$) and *alternating on purple* ($\beta = 0.06, \text{CrI} = [-0.03, 0.15]$) norms. Though the frequency of *stable on orange* in this condition was reduced relative to the control, the plurality of participants converged on this strategy, suggesting that its optimality with respect to fairness and coordination cost outweighted the downside of a non-optimal joint reward and the complexity of *alternating on purple*.

In Condition 3 (Fairness vs. Complexity), participants developed a norm of *stable selection on purple* more often, compared to the control ($\beta = 0.48, \text{CrI} = [0.32, 0.64]$) and to Condition 2 ($\beta = 0.35, \text{CrI} = [0.17, 0.53]$). They also developed an alternating norm more frequently than in the control ($\beta = 0.10, \text{CrI} = [0.01, 0.20]$). Correspondingly, they were less likely to converge on *stable selection on orange* compared to the control ($\beta = -0.75, \text{CrI} = [-0.92, -0.59]$) and to Condition 2 ($\beta = -0.29, \text{CrI} = [-0.46, -0.13]$), and failed to form any norm more frequently compared to the control ($\beta = 0.18, \text{CrI} = [0.04, 0.31]$). The increased prevalence of the *stable on purple* norm in Condition 3 compared to Condition 2 suggests that participants' aversion to unfair outcomes was graded: they converged more frequently on an unfair norm when the price discrepancy was lesser. In both Conditions 2 and 3, alternation occurred relatively infrequently, suggesting that this more complex strategy was 1) more difficult for participants to initially conceive, or 2) more difficult to successfully implement.

Model results

Alternative models We compared our model against 2 ablated models which used just one of the reward or complexity terms in the utility calculation. The first alternative model is the reward-only model, which is equivalent to setting w_c to 0 in the full model. The second alternative is the cost-only model, which is equivalent to setting w_j, w_f both to 0.

Parameter estimation We adapted the general formulation of this family of models to our specific task and fit the parameters of the model to the experimental data. The model represents candidate joint plan as states in the joint state space. For example, the state $[(o_1, o_2), (o_1, o_2), (o_1, o_2), (o_1, o_2)]$ represents the norm of player 1 selecting the 1st orange spot and player 2 selecting the second orange spot. We assumed a finite horizon of $t = 4$ trials, which made finding the optimal value function via policy iteration computationally tractable, but there is no distinction between a t of 4 or 100: all states in which one player selects the 1st orange and the other selects the 2nd orange at each timestep represent the same joint plan.

For every candidate joint plan, we computed the joint reward, fairness and coordination cost. Each of those 3 terms has an associated free parameter which represents the relative weighting of that component within the overall utility. To make weight parameters more straightforward to interpret, we normalized each of the joint reward, fairness and cost values to a value between 0 and 1. For joint reward and fairness, this is done by linearly rescaling the optimal outcome for the given parking lot (i.e., the smallest possible joint payment and the smallest possible gap in cumulative amount paid between the two players) to a value of 1, and the least optimal outcome to 0. We set $w_c = \frac{1}{3}$ for the complexity penalty.[3]

After fixing w_c, the model includes 3 free parameters: two weight parameters w_j, w_f, and the softmax optimality parameter β. To approximately fit model parameters to our data, we performed a grid search over the following ranges: $\beta \in [1, 2, 3, ..., 15]; w_j, w_f \in [0, 0.11, 0.22, ..., 1.98]$. To quantify model fit to the data we computed the Jensen-Shannon divergence (JSD) between the predicted and the empirical distribution of norm types for the 3 conditions. To account for potential overfitting, we split the data into train and test sets with a 70-30 split and selected the best parameter values based on minimum mean JSD across the 3 conditions, using data from pairs in the train set only. The metrics reported here reflect the mean JSD values on the unseen test set.

Model predictions For each of the 3 models, we computed the probability of selecting each type of norm under all combination of parameters.[4] The best-fitting full model (parameter values $\beta = 13, w_j = 1.43, w_f = 0.22$) closely fit the human data (mean JSD=0.14), far outperforming the best-fitting reward-only model (mean JSD=0.63) and the best-fitting cost-only model (mean JSD=0.57). Figure 3 shows the predicted distribution of norm categories under the best-fitting parameterization of the full model, plotted against the empirical distribution for all 3 conditions.

The inferred value of $w_j = 1.43$ for the best-fitting model was substantially higher than $w_f = 0.22$ and (fixed) $w_c = 0.33$, providing quantitative evidence that people weigh jointly-optimal reward more heavily than a fair outcome or reducing complexity when generating norms, consistent with

[3] This is because the longest possible program generating a joint plan of length 4 is 3 operations, so the most complex joint plan would have an associated cost of $w_c \cdot C(\tau) = \frac{1}{3} \cdot 3 = 1$.

[4] The cost-only model has fewer free parameters, so for this model we only searched over the value of β.

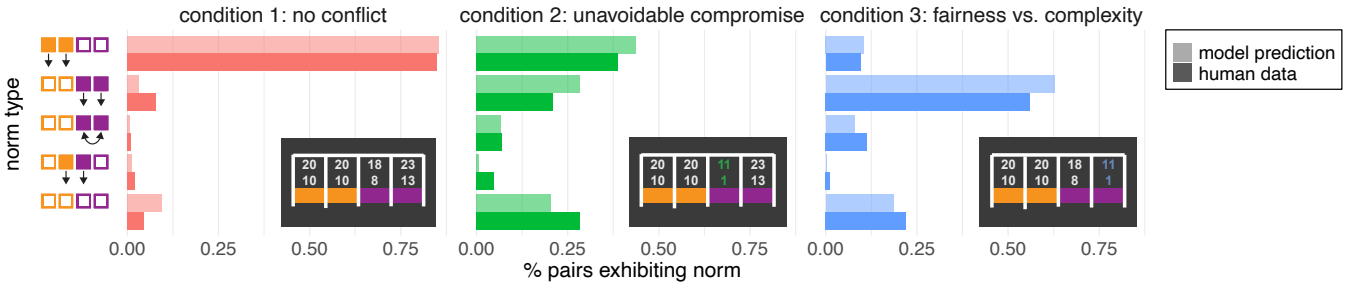

Figure 3: Full model predictions and empirical distributions of norms across the 3 conditions. Parking lot reward structure for the condition is shown in inset. Empirical distribution is based on data from all pairs.

the qualitative patterns observed in Conditions 2 and 3.

Though our model overall fit the data well, it was least aligned with participant behavior in Condition 2. The higher rate of failure to form any norm in this condition suggests there was something particularly difficult about this reward configuration. One possible explanation is that because people's preferences for different norms were less concentrated, there was a lower probability that both people in a pair were aligned on the same joint plan, resulting in a higher probability of coordination failure.

To assess the whether the fit of our model to the behavioral data reflects an overly expressive model class rather than a well-aligned theory, we analyzed the model's capacity to fit randomly-generated datasets with the same structure (triplets of 5-category probability distributions). If the model better fits our behavioral data compared to randomly-generated datasets, this provides evidence that the model is capturing something meaningful about the process generating the data; in contrast, if the model can fit any distribution as well as the experimental data, its potential to offer insight is more limited. We generated 1000 null datasets and computed the mean JSD between each null dataset and the predictions of the model under the parameter values that led to the best fit on that null dataset. The JSD distribution from the null datasets has a mean of 0.32 and standard deviation of 0.05. In contrast, the best full model fit to the experimental data achieves a mean JSD of 0.14 (falling in the first percentile of the null dataset distribution). This result provides evidence against an overly-expressive model class and indicates alignment between this model's dynamics and participant behavior.

Exploratory analysis of written plans

Before starting the game, participants wrote strategic plans. We conducted an exploratory analysis of the plans to better understand the relationship between participants' individual intentions and the norms that emerged over the course of interacting with their partners. Following past work (Gilardi, Alizadeh, & Kubli, 2023; Rathje et al., 2024), we used a large language model (GPT-4) to classify written plans across conditions according to the category of norm expressed, using the same 5-category classification scheme as previous analyses.

We evaluated the extent to which 0, 1 or 2 participants in a pair writing a plan that described a particular strategy predicted implementation of that strategy during the game. The number of players expressing a plan was predictive of plan implementation for the three major types of norms: *stable on orange* ($\beta = 0.28, \text{CrI} = [0.15, 0.40]$), *stable on purple* ($\beta = 0.27, \text{CrI} = [0.08, 0.45]$) and *alternating on purple* ($\beta = 0.28, \text{CrI} = [0.19, 0.38]$). These results suggest that the initial plans that people conceptualized did influence the types of norms that developed through interaction.

Discussion

The ability to form complex norms is underpinned by complex social reasoning. Accordingly, the core cognitive principles involved in norm formation have been difficult to capture with traditional models. We developed an account of norm formation that views this process as a form of joint planning in which participants trade off joint optimality, fairness and complexity. We formalized this theory using an integrative computational model designed to capture aspects of the process by which people simulate joint plans. The model was designed to express an overall preference for simpler joint plans, but to be flexible enough to account for adaptive formation of more complex strategies such as alternating in situations that demand additional structure. The model's predictions closely aligned with the distribution of norms participants formed in an iterated decision-making task.

Our model is limited in important ways. One key limitation is that it does not account for individual learning over the course of an interaction and does not make predictions about how an individual behaves conditioned on a history of interactions. Moving forward we hope to investigate how participants adapt their strategies based on their partners' actions via inverse planning (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Jara-Ettinger, Schulz, & Tenenbaum, 2020): observing a sequence of actions and inferring the other player's intentions that could have given rise to that behavior.

Here we focused on characterizing the cognitive principles that enable norm formation by studying behavior in dyads, as opposed to in larger groups or communities. Understanding how strategies learned in one-on-one interactions ultimately give rise to societal-scale norms (C. M. Wu, Dale, & Hawkins, 2023) is an important avenue for future work.

Acknowledgments

We thank Jevan Yu and the anonymous reviewers for helpful suggestions.

References

Axelrod, R. M. (1984). *The evolution of cooperation* (Rev. ed ed.). New York: Basic Books.

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), 1–10.

Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms* (1st ed.). Cambridge University Press.

Gavrilets, S., Tverskoi, D., & Sánchez, A. (2024). Modelling social norms: an integration of the norm-utility approach with beliefs dynamics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *379*(1897), 20230027.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023, July). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30).

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science*, *32*(1), 108–154.

Hawkins, R. X. D., Goodman, N. D., & Goldstone, R. L. (2019). The Emergence of Social Norms and Conventions. *Trends in Cognitive Sciences*, *23*(2), 158–169.

Helbing, D., Schönhof, M., Stark, H.-U., & Hołyst, J. A. (2005). How individuals learn to take turns: emergence of alternating cooperation in a congestion game and the prisoner's dilemma. *Advances in Complex Systems*, *08*(01), 87–116.

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

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

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction. In *Proceedings of the Cognitive Science Society.*

Kleiman-Weiner, M., Sosa, F., Thompson, B., van Opheusden, B., Griffiths, T. L., Gershman, S., & Cushman, F. (2020). Downloading Culture.zip: Social learning by program induction. In *Proceedings of the Cognitive Science Society.*

Le Pargneux, A., Chater, N., & Zeitoun, H. (2023, May). *Contractualist tendencies and reasoning in moral judgment and decision making.* PsyArXiv. Retrieved from `osf.io/preprints/psyarxiv/p4cyx`

Levine, S., Chater, N., Tenenbaum, J., & Cushman, F. (2023). *Resource-rational contractualism: A triple theory of moral cognition.* Retrieved from `https://osf.io/p48t7`

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning* (pp. 157–163).

Misyak, J. B., & Chater, N. (2014). Virtual bargaining: a theory of social decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012, May). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199–217.

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C., & Bavel, J. J. V. (2024). *GPT is an effective tool for multilingual psychological text analysis.* Retrieved from `https://osf.io/sekf5`

Thielmann, I., Böhm, R., Ott, M., & Hilbig, B. E. (2021, February). Economic Games: An Introduction and Guide for Research. *Collabra: Psychology*, *7*(1), 19004.

Tomasello, M. (2019). *Becoming Human: A Theory of Ontogeny*. Harvard University Press.

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.

Wu, C. M., Dale, R., & Hawkins, R. D. (2023). *Group coordination catalyzes individual and cultural intelligence.* Retrieved from `https://osf.io/gscy6` (Publisher: OSF)

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.