Balancing Competing Goals when Coordinating over Rule Construction

Jess Mankewitz (jmank@stanford.edu)[*1,2] **Katherine M. Collins (katiemc@mit.edu)**[*3,4,5]
Joshua B. Tenenbaum[3] **Thomas L. Griffiths**[4] **Robert D. Hawkins**[1]

[1]Stanford University [2]UW–Madison [3]MIT [4]Princeton [5]University of Cambridge
[*]Equal contribution

Abstract

Even competitive activities require cooperation to agree on the rules. How do people navigate this tension, and what conditions help them reach agreement? We study rule negotiation in a collaborative game design task ($N = 184$) where dyads propose and vote on what game to play, crossing two factors: communication (chat available vs. vote-only) and role uncertainty (whether players know which role they will occupy). We find communication increased agreement rates and role uncertainty had only a minimal impact on agreement. When players knew their role, they proposed games that subtly favored themselves, balancing self-interest against the need for acceptance. Chat content was largely coordinative rather than argumentative: players declared preferences and deferred rather than playing hardball. Our findings highlight how deliberation helps people coordinate on rules, even when they know which side they're on.

Keywords: coordination; gameplay; institution formation; game creation; resource-rationality

Introduction

Even competitive activities may require cooperation to get off the ground. Before legislators can debate a bill, they must agree on what reaches the floor. Before players can compete in a game, they must agree on the rules. In each case, people jointly construct the arena within which they'll act, but their interests in what that system looks like may conflict. This requires not just reasoning about what a system will be like, but reasoning about what others will be likely to accept. How do people navigate this tension?

Political theorists have offered frameworks for thinking about this problem. For example, Ostrom distinguished three nested levels of rules: *operational rules* governing day-to-day decisions, *collective-choice rules* through which participants can modify operational rules, and *constitutional rules* that determine who participates in collective choice and how. Different considerations are relevant at each level. Operational choice asks "given these rules, what maximizes my payoff?" Collective choice asks "what rules would we all be willing to live under?" The latter requires a kind of mutual acceptability that the former does not. Ostrom argued this is achieved through deliberative forums, settings where affected parties can communicate about and jointly modify the rules governing their interactions.

But how is mutual acceptability achieved when interests conflict? Rawls proposed one solution: if choosers do not know their position in the resulting system (a "veil of ignorance") self-interest and impartiality align. Favoring one role risks disadvantaging yourself. Empirical work has validated that veil-of-ignorance reasoning leads to choices favoring the greater good (Huang, Greene, & Bazerman, 2019), and that these effects reflect genuine impartiality. Laboratory studies of constitutional choice similarly find that when participants select distributive principles without knowing their future position, they converge on hybrid principles that combine efficiency with a floor constraint on the worst-off (Frohlich & Oppenheimer, 1992).

There are two major gaps in this literature. First, most empirical work uses one-shot individual decisions, while real constitutional design is iterative and deliberative: parties propose, disagree, communicate, and (sometimes) converge. Though Ostrom emphasized the role of deliberative forums, empirical work on the veil of ignorance has largely examined isolated choosers. Second, cooperation depends on shared expectations of fair treatment (Rand et al., 2013; Yu & Thompson, 2024), and people enforce fairness norms even at personal cost (Kahneman, Knetsch, & Thaler, 1986). But this work has focused on behavior within fixed rules (rejecting unfair offers, punishing defectors) rather than on how people negotiate the rules themselves. And, whether people even negotiate at all—or engage in other communicative strategies.

Games provide an ideal domain for studying constitutional choice at a tractable scale. They have clear rule systems, quantifiable payoff structures, and a natural balance of competitive and cooperative elements. Recent work has characterized how individuals evaluate novel games through bounded probabilistic simulation (Collins et al., 2025a) and how they create new games (Collins et al., 2025b). Their Intuitive Gamer model well-captures how people (before ever playing a game) judge whether a game is fair or fun. However, in both works, game evaluation and creation are done by a single reasoner, and while gameplay is with another person, players have no choice over what they will play. However, with games—as in many other areas of life—decisions about what game to play require coordination with other people. When two players must agree on rules, their preferences may conflict: each might prefer rules that favor themselves, but not so much that the other refuses to play.

We study how dyads decide what competitive game to play, across two factors central to constitutional theory:

- *Communication*: Can participants only vote on proposals, or can they deliberate before deciding?
- *Role uncertainty*: Do participants know which player they will be, or must they design rules from behind a veil?

These manipulations yield distinct predictions: (1) Communication should facilitate deliberation about the rules.

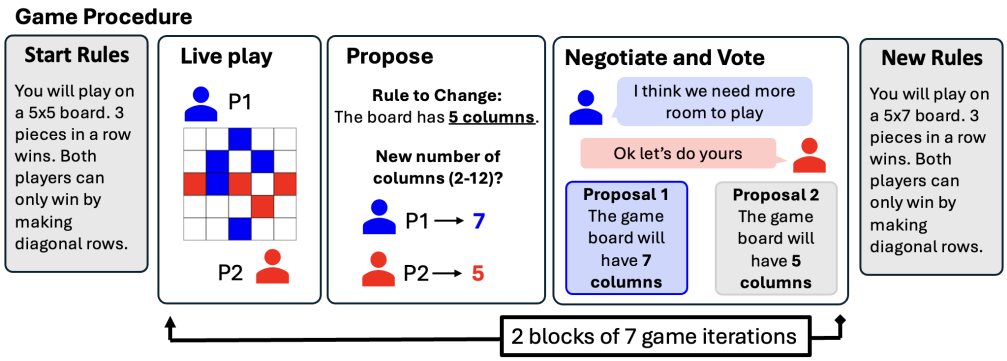

Figure 1: **Procedure.** Players were informed of a game, then play one round of that game. After play, they began a series of proposal-vote rounds. At each round, they were told a rule that they can propose to change (or stay with the current value). In the chat condition, players could communicate over said rule choice. They then played the resulting game and began another proposal-vote round.

Rather than each person independently simulating what the other would find acceptable, they can simply ask. This should accelerate convergence. It is an open question *what kind* of communication such a channel would permit: do players negotiate, haggle, and bargain; or does the channel simply help players coordinate on whose preferences to adopt. Indeed, experimental work on cheap talk in coordination games consistently finds that communication improves coordination on efficient equilibria (Cooper et al., 1992; Charness, 2000; Charness & Grosskopf, 2004), though this work has focused on strategic choice within fixed games rather than collective choice about the games themselves. And (2) the veil of ignorance should promote fairer game structures by aligning self-interest with impartiality: when participants do not know whether they will be Player 1 or Player 2, proposing rules that advantage one role risks disadvantaging themselves. However, when participants do know their role, even though they may prefer games that favor their role—deciding on what game to play requires a higher-level of cooperativity that may dissuade a reasoner from attempting to exploit knowledge of their role to advantage themselves.

Our work makes three contributions. First, we extend the experimental settings of solo game evaluation and game creation from (Collins et al., 2025a, 2025b) toward *joint* game design, asking how dyads coordinate on what to play. We do so in a subspace of the two-player competitive strategic board games, wherein individual reasoning has been characterized in their prior work. Second, we provide a paradigm for studying iterated constitutional choice at a tractable scale, moving beyond the one-shot distributive decisions that have dominated this literature. Third, we examine how language facilitates (or fails to facilitate) the construction of shared understandings of fairness—and what kinds of communication strategies people employ—connecting political economy to work on common ground and collaborative cognition.

Methods

We first describe the game negotiation paradigm that we develop, before detailing our stimuli, procedure, participant pool, and analysis methodology.

Game negotiation domain Unlike in Collins et al., wherein people play one round of one game in a pre-specified game, here, players have the option to choose what game to play next.

Players play one round of a given game and are then told one feature or rule of the game that they can change (either the number of board rows or columns; the number of pieces Player 1 and/or 2 need in a row to win; and whether there are any special conditions on diagonal wins, e.g., only diagonal wins allowed, Player 2 cannot win diagonally, etc.). Each player suggests a value for that rule feature. In the veiled condition, players do not know their role (as Player 1 or 2) prior to making their suggestion; in the unveiled condition, they do know. Players make their suggestions simultaneously.

After both players make their suggestion, they are told of the other players' suggestion. If both players made the same suggestion, they immediately play that game. If they proposed different values, they then must vote on which rule value they would like to play. In the chat condition, players can communicate before casting their vote. If they do not agree, then they are forced to play with random value from the options for that feature (See Figure 1 for a schematic).

Players then play a round of that new game; after playing, they proceed with another round of proposal and agreement.

Procedure We ran a 2×2 design: participants either did or didn't know their role (unveiled vs. veiled) before making the proposal for the next match and where participants do or don't have access to chat before voting on said proposal. Participants were randomly allocated to one conditions.

In each condition, players play seven rounds off of a single base game, where they have the option to change one rule at

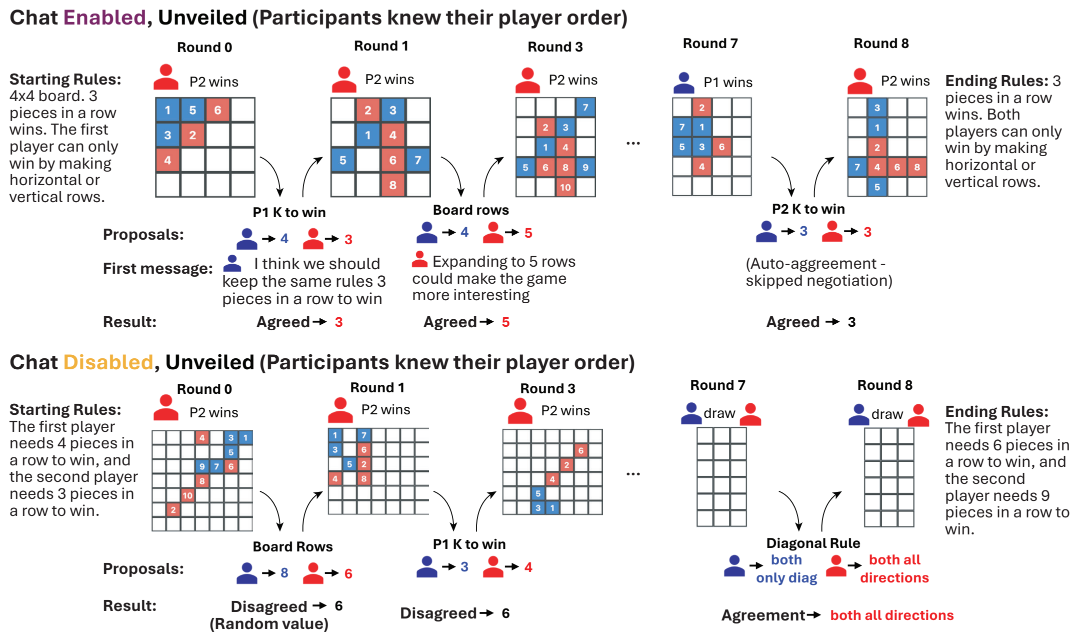

Figure 2: **Example game trajectories.** Two representative dyads showing sequences of gameplay, game rule proposal, voting, and where applicable chat (top row). Agreement indicates that players decided to vote on the same game, even if they proposed different games initially. When players vote on different games, they are required to play a random game.

a time (that rule is given to them at each round). Participants play from two base games, totaling 14 individual matches. The setting is similar to a series of two short Gibbs Sampling with People chains (Harrison et al., 2020). We chose this number of rounds and base games to balance long enough chains to being to assess convergence behavior, across a diversity of games, while aiming to avoid participant fatigue.

We extend the human-human play interface from Collins et al. built on Empirica (Almaatouq et al., 2021) to have a proposal and voting phase; as in the prior work, during the match, participants have the option to request a draw or surrender if they want to end the game early. Participants were told we would select a random match from block and they would receive a bonus of $1.50 USD if they won the match.

Our primary measure is whether players agree on a game. We secondarily look at the choice and estimated payoff of these games, and amount and type of chat used (in the chat condition).

Participants We recruited 200 English-speaking participants from the online data collection platform Prolific (Palan & Schitter, 2018). Participants took on average 31.5 minutes to complete the game and were compensated an average of $13 USD. We excluded 8 dyads (16 participants) due to inactivity and retained a final sample of 184 participants.

Stimuli Participants play and select games from the 121 subspace of two-player perfect information competitive board games from Collins et al.. Games are variations off of more familiar games like Tic-Tac-Toe, where players take turns placing pieces on an $M \times N$ board and attempt to make K pieces in a row. Games may have special rules, e.g., where Player 1 and 2 need to make different K in a row to win, or where one or more players cannot win diagonally. We choose 7 games which have a bias towards a player or are likely to end in a draw as seeds; we choose these games to encourage proposals away from them.

Game analysis The games are all expressible in a symbolic game description language (DSL). Here, we use the board game-optimized DSL, Ludax (Todd et al., 2025). To estimate the fairness of the games created, we ran the Intuitive Gamer model (Collins et al., 2025a) to capture how boundedly-rational novice reasoners may estimate the expected outcomes of new games. However, we note that in this work, we use these models as analysis tools; the Intuitive Gamer was focused on modeling novice play in the *first* round of play (or even *before* the first round). Here, players may play multiple rounds of the same game. Future work ought to look at what kind of game reasoning people are engaging in when deciding what to play together, especially when they may play the same game or only slight variations repeatedly.

Text processing We coded the language used by participants in the chat condition using an LLM-assisted pipeline. Each message was annotated based on (a) whether it invoked

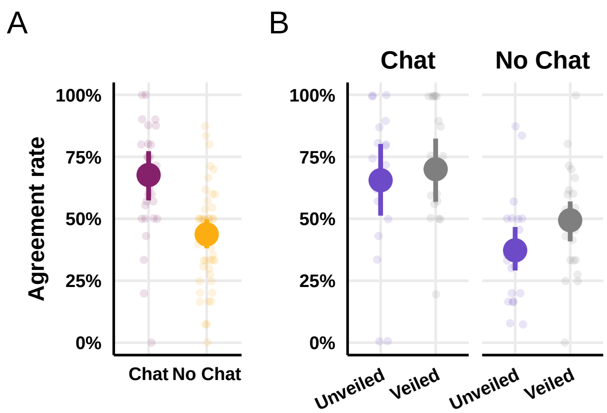

Figure 3: **Agreement rates.** **(A)** Average agreement rates over what game to play when a vote was necessary (i.e., when participants did not both propose the same game). The chat condition includes only rounds where players exchanged at least one message. **(B)** Average agreement for games where voting was necessary, broken down by chat access and role uncertainty. As in **(A)**, the chat conditions include those where at least one chat exchange occurred. Error bars are bootstrapped 95% CIs around the mean.

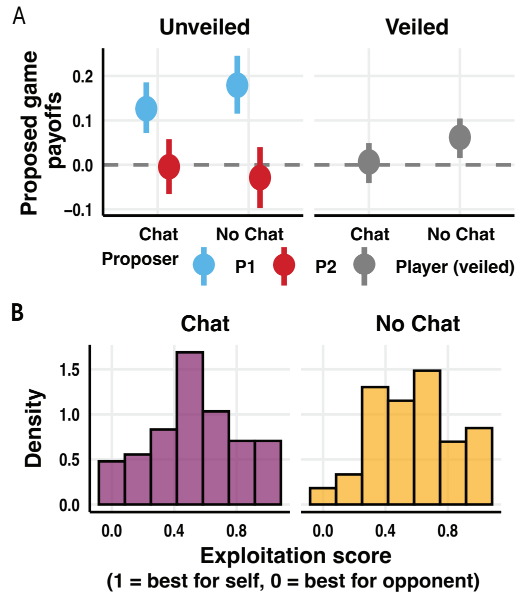

Figure 4: **Expected advantage of proposals.** **(A)** Average expected game payoff (under Intuitive Gamer simulations) made by players in each condition. Payoff is relative to the first player (P1). Higher means greater expected advantage to P1; negative values mean P2 has an expected advantage. Most possible games have an advantage towards P1. **(B)** Games where players decide on K provide a direct picture into how exploitative players are in their proposals. A player should want to minimize the K they need to win and maximize the K the other player needs to win. We define a player proposals' level of exploitation as how far it gets to perfectly achieving that objective (based on the board size); a score of 0 means the player proposed a game that would completely favor their opponent (e.g., low K for opponent). Error bars are bootstrapped 95% CIs around the mean.

a rule *quality* as a justification for a proposal—fairness, fun, efficiency, reciprocity, or other—and (b) the *referent* of that quality claim: whether it applied to the speaker (self), their partner, a specific player position (Player 1 or Player 2), or the game in general. We counted only explicit justifications (e.g., "more fair," "drags the game out"), not bare preferences (e.g., "let's do 5") or simple coordination (e.g., "ok," "I'm with you"). A language model (Gemma-2 9B; (Gemma Team, 2024)) was prompted to annotate each message with its full within-round conversation context and current game state. An author reviewed all language model annotations to verify the prompt was capturing the intended distinctions. The processed dataset contains 748 messages across 227 rounds with chat; 23% of messages contained at least one justification.

Results

Access to chat helps agreement Participants in the chat condition agreed over what game to play significantly more than participants without access to chat (Figure 3a). A generalized linear mixed-effects model with random intercepts for dyad confirmed that access to chat was associated with 2.56 times higher odds of agreement (95% CI [1.64, 3.99], $p < .001$). In 35.6% of rounds, players proposed the same value and agreed immediately; among the remaining vote-needed rounds, no-chat dyads reached agreement 42.5% of the time, compared to 58.6% in the chat condition. However, these analyses are correlational; dyads differed in their chat usage: of the instances where players made different proposals (57% of the rounds in the chat-enabled condition*), 63.4% of those rounds included at least some kind of chat exchange. We fit an additional generalized linear mixed model restricted to the chat condition and confirmed that rounds where players used chat had higher odds of agreement than rounds where they did not (OR = 4.17, 95% CI [2.12, 8.20], p $<$ .001).

Players take advantage of known role, but with minimal impact on coordination There was no statistically significant effect of the veil on agreement ($p = .77$), and this effect did not differ between chat and no-chat conditions ($p = .37$; Figure 3b). This is notable: players who knew their role had the opportunity to propose rule settings which substantially advantage this role, or, as the games are zero-sum, actively hinder their opponent. One may therefore expect that role knowledge would lead to systematically lower agreement rates. But this is not what we see. The minimal impact on co-

*70.8% of the no-chat condition games did not involve auto-agreement.

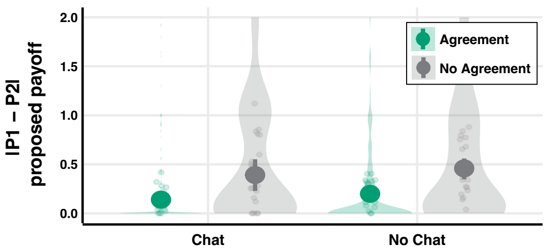

Figure 5: **Proposal bias and agreement rate.** Average payoff differential (absolute difference between P1 and P2 payoffs) for players with and without chat, for cases where they agree versus do not find agreement.

ordination despite unveiling roles appears to be from players proposing games that often lean in their favor — but not very far. Assessing the expected payoff of players' proposals under Intuitive Gamer simulations reveals that players in the unveiled condition, tended to propose games to their advantage (Figure 4a), particularly the first player (which may be due to the space of games generally favoring the first player). A linear mixed model confirmed this role-based bias in proposed payoffs ($\beta = 0.065$, $p = .002$), which was attenuated under role uncertainty (veil $\times$ proposer interaction $p = 0.08$). Yet, these proposals are typically only slightly biased in a player's own favor (Figure 4b). These proposals may be an attempt to balance competing goals of trying to win, but also not propose a game that is so biased the other player would never vote for it (or may get insulated for later round proposals). When proposals are substantially biased, then players do have higher rates of disagreement. In cases of failed agreement, this was more often when there was a higher difference in the expected payoff (bias) between players: for each 1-unit increase in payoff differential, the odds of agreement decreased by 36% (OR $= 0.64$, 95% CI $[0.47, 0.88]$, $p = .006$; Figure 5).

Players coordinate, not negotiate In exploratory analyses, we examined the content of the 748 messages sent across rounds where players communicated. Most chat exchanges were brief (6.6 words per message on average; and 3.3 messages per rounds on average): players typically declared their preferences or deferred to their partner rather than engaging in extended negotiation (Table 1). When players did invoke explicit justifications for their proposals, fairness was the most common quality mentioned (98 messages), followed by fun (43), and efficiency (23) (Fig 6). These justifications most often referred to the game in general (e.g., "that would be fairer") rather than to a specific player's payoff: approximately 79% of quality mentions used a general referent, while 13% were self-referent (e.g., "impossible for me to win").

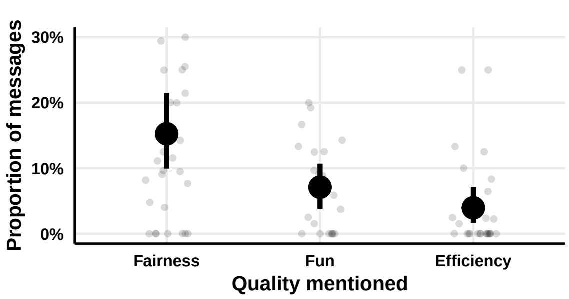

Figure 6: **Descriptive chat analyses.** Proportion of messages sent by each dyad, explicitly referencing a particular game quality as their justification for a rule change.

Agreed	**Disagreed**
Initial: 9×3 board *Rule:* board columns Player A → 2, Player B → 3 *Final:* 2 (agreed)	*Initial:* 10×10 board *Rule:* board columns Player A → 5, Player B → 10 *Final:* 10 (random)
B: I'll go with you this time. **A:** ok	**A:** hey lets make it 5
Initial: P2 needs 3 *Rule:* P2 pieces to win Player A → 3, Player B → 4 *Final:* 3 (agreed)	*Initial:* win all directions *Rule:* diagonal settings Player A → all, Player B → P2-only *Final:* P1 no diagonal (random)
A: Three pieces to win isn't fair. **B:** Impossible to win with 4 in a row **A:** Only fair for player who starts first **B:** it'll always end a draw **A:** How will you vote? **B:** 3 **A:** Fine	**B:** I think we screwed up choosing different amounts that we need to win **B:** so to even it out - you need 3 but can only be diagonally, and I need 5 but in any direction **B:** I think this is what they are trying to get us to agree to. **A:** Any direction is all I will go with. *[4 lines of chit-chat omitted]* **B:** 5 v 3 is unfair **B:** so if you need 3 but diagonally, it evens it out **A:** I know that, I am easy, so whatever **B:** choose my proposal then :)

Table 1: **Example chat exchanges from four negotiation rounds.** All examples are from the veiled condition, where players did not know which player position (P1 or P2) they would occupy. Players are labeled A and B rather than P1/P2 to reflect this anonymity. References to player positions within the chat (e.g., "player who starts first") reflect participants reasoning about the game rules in the abstract, not knowledge of their own role.

The amount of chat did not differ significantly by veil condition ($\beta = 4.36$, $p = .32$), nor did the types of qualities or referents invoked (all $p > .05$). Coordination via chat appears to have been accomplished primarily through simple preference declaration and mutual accommodation rather than through elaborate negotiation.

Language as a tool for meta-level coordination Thus far, we have focused on the role language can play in helping people coordinate over what game to play; however, language can help people coordinate over all kinds of games, including how to game the system in which they are engag-

Round 1	Round 2
Initial: 4×10 board, P1:6 P2:10, all directions *Rule:* P2 pieces to win P1: 1, P2: 5 Final: 1 (agreed)	*Initial:* 10×10 board, P1:9 P2:9, all directions *Rule:* diagonal settings P1: P1-only, P2: both-only Final: P1-only (agreed)
P1: One of us has to let the other win **P2:** ok **P1:** so who wins?... **P2:** we pick yours **P1:** ok	**P2:** this is a silly game lol **P1:** I want out! I'll let you win **P2:** ok i pick yours **P1:** ok

Table 2: **Example collusion.** Rounds where players used chat to coordinate who would win rather than negotiate the rule.

ing. Our study was designed to test how players decide what rules of the game to play. This is mechanism design in a microcosm (Hurwicz, 1973; Maskin, 2008). Yet, we as experimenters are also playing a game of mechanism design—designing the rules by which players can decide what to play. In pilots, we regularly observed that without careful incentives, players would regularly use language to coordinate over ways to "game" our game, e.g., deciding they would trade off who would win. Even in our current design, we observed a few traces of collusion in some chats (e.g., Table 2). Future work can better explore the interplay between experimental mechanism design and players' own mechanism design.

Discussion

In general, access to a communication channel aids coordination over what rules to play by. Unveiling players' roles has minimal impairment on coordination. Even though players tend to capitalize on that prior knowledge to propose rulesets that favor their role—they only do so subtly. This hints at a subtle dance players are playing; when they know their role, as the games are zero-sum, they tend to propose games that give them a slight advantage, yet, they are also seemingly sensitive to the cooperative aspects of rule agreement (a player is often not going to agree to a game that is completely biased against them). The way players communicate to reach agreement is largely coordinative rather than argumentative: players typically declared preferences and deferred to one another rather than engaging in extended negotiation. When explicit justifications were offered, they most often appealed to fairness or fun in a general sense. Role uncertainty did not significantly change either the amount or content of chat.

Our findings connect to recent theoretical work on how people navigate social decisions with conflicting interests. The resource-rational contractualism framework (Levine et al., 2024a, 2024b) suggests that moral cognition is organized around approximations of an idealized contractualist procedure, gathering affected parties to negotiate mutually acceptable agreements. Because this idealized procedure is intractable, people engage in *virtual bargaining*, mentally simulating what agreement would be reached if negotiation could hypothetically occur. Like the Intuitive Gamer model of game evaluation, virtual bargaining trades off accuracy for tractability in resource-rational ways. But our work here suggests that at least in this case, people may not be explicitly negotiating, but rather "hedging" their proposals (when their role is unveiled) to not substantially favor their own role, or are otherwise declaring their preferences in chat to coordinate not to haggle.

Our paradigm offers a way to investigate what happens when people have the option to engage in a real iterative coordination, instead of only simulating that process. When participants can freely discuss, computational demands shifted, as now, another player can be queried for their preferences to offload computation. This contrast raises several questions that our paradigm is well-suited to address in future work. Does real deliberation achieve outcomes that virtual bargaining alone cannot? If communication allows people to "offload" the computational work of simulating others' preferences, we might expect faster convergence and perhaps qualitatively different solutions—integrative modifications that benefit both players in ways that are difficult to discover through independent mental simulation.

While in this particular mechanism, we find that enabling people to communicate via language to coordinate on a solution, there are other mechanisms that one could design where language may be less valuable, e.g., if players can iteratively propose and accept or reject suggestions (like Markov Chain Monte Carlo with People (Sanborn & Griffiths, 2007)). Future work could better understand the critically of *linguistic* communication, as opposed to other communication, in assessing the extent to which any communication channel helps people balance competing goals in rule construction. In ongoing work, we are also looking at how the number of rounds (here, always 7) impacts coordination behavior. Additionally, participants were always told which rule they can change. We may expect differences in coordination efficiency if players can also choose which rule (and how many rules) to modify at any one time. Lastly, while we focus here on conflicts between player advantage and agreement, there may be other factors players consider when deciding what to play (e.g., whether both players will find a game fun). Future work can expand the assessments of how people coordinate over a wider range of multi-dimensional goals.

More broadly, our paradigm opens up questions about collaborative mechanism design in more complex systems. Whenever groups must jointly decide on the rules governing their interactions (from workplace policies to platform governance to international agreements) similar dynamics of constitutional choice under uncertainty and communication may apply. But we still understand relatively little about the cognitive processes underlying successful institutional design at any scale. This may be especially important as AI systems may be increasingly used in part of the mechanism design process (Koster et al., 2022). Games offer a uniquely tractable window into these processes (Frey et al., 2023). They have clear rules, quantifiable payoffs, and admit computational models of evaluation and play. By studying how people negotiate the rules of games, we can begin to characterize the cognitive capacities that underlie this distinctively human ability to jointly construct the strategic environments we inhabit.

Acknowledgments

We thank Lio Wong, Graham Todd, Ced Zhang, Max Kleiman-Weiner, Sydney Levine, and Scott Stern for conversations that informed this work. KMC acknowledges support from the NSF SBE SPRF.

References

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021). Empirica: a virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*, *53*(5), 2158–2171.

Charness, G. (2000). Self-serving cheap talk: A test of aumann's conjecture. *Games and Economic Behavior*, *33*(2), 177–194.

Charness, G., & Grosskopf, B. (2004). What makes cheap talk effective? experimental evidence. *Economics Letters*, *83*(3), 383–389.

Collins, K. M., Todd, G., Zhang, C. E., Weller, A., Togelius, J., Chu, J., ... Tenenbaum, J. B. (2025b). Generation and evaluation in the human invention process through the lens of game design. *arXiv preprint arXiv:2508.10914*.

Collins, K. M., Zhang, C. E., Wong, L., Barba, M., Todd, G., Weller, A., ... Tenenbaum, J. B. (2025a). People use fast, flat goal-directed simulation to reason about novel problems. *arXiv preprint arXiv:2510.11503*.

Cooper, R., DeJong, D. V., Forsythe, R., & Ross, T. W. (1992). Communication in coordination games. *The Quarterly Journal of Economics*, *107*(2), 739–771.

Frey, S., Hedges, J., Tan, J., & Zahn, P. (2023). Composing games into complex institutions. *Plos one*, *18*(3), e0283361.

Frohlich, N., & Oppenheimer, J. A. (1992). *Choosing justice: An experimental approach to ethical theory*. University of California Press.

Gemma Team, G. D. (2024). *Gemma 2: Improving open language models at a practical size.*

Harrison, P., Marjieh, R., Adolfi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., ... Jacoby, N. (2020). Gibbs sampling with people. *Advances in neural information processing systems*, *33*, 10659–10671.

Huang, K., Greene, J. D., & Bazerman, M. (2019). Veil-of-ignorance reasoning favors the greater good. *Proceedings of the National Academy of Sciences*, *116*(48), 23989–23995.

Hurwicz, L. (1973). The design of mechanisms for resource allocation. *The American Economic Review*, *63*(2), 1–30.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the assumptions of economics. *Journal of business*, S285–S300.

Koster, R., Balaguer, J., Tacchetti, A., Weinstein, A., Zhu, T., Hauser, O., ... others (2022). Human-centred mechanism design with democratic ai. *Nature Human Behaviour*, *6*(10), 1398–1407.

Levine, S., Chater, N., Tenenbaum, J. B., & Cushman, F. (2024b). Resource-rational contractualism: A triple theory of moral cognition. *Behavioral and Brain Sciences*, 1–38.

Levine, S., Kleiman-Weiner, M., Chater, N., Cushman, F., & Tenenbaum, J. B. (2024a). When rules are over-ruled: Virtual bargaining as a contractualist method of moral judgment. *Cognition*, *250*, 105790.

Maskin, E. S. (2008). Mechanism design: How to implement social goals. *American Economic Review*, *98*(3), 567–576.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.

Palan, S., & Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Rand, D. G., Tarnita, C. E., Ohtsuki, H., & Nowak, M. A. (2013). Evolution of fairness in the one-shot anonymous ultimatum game. *Proceedings of the National Academy of Sciences*, *110*(7), 2581–2586.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Sanborn, A., & Griffiths, T. (2007). Markov chain monte carlo with people. *Advances in neural information processing systems*, *20*.

Todd, G., Padula, A. G., Soemers, D. J., & Togelius, J. (2025). Ludax: A gpu-accelerated domain specific language for board games. *arXiv preprint arXiv:2506.22609*.

Yu, D., & Thompson, B. (2024). People balance joint reward, fairness and complexity to develop social norms in a two-player game. In *Proceedings of the 46th annual meeting of the cognitive science society.*