The Content of Common Ground Shapes Coordination

Yuka Machino (ymachino@stanford.edu)[1] **& Robert Hawkins**[2]
[1]Department of Computer Science, Stanford University
[2]Department of Linguistics, Stanford University

Abstract

Common ground is widely understood to facilitate coordination, but what happens when the shared expectation is competitive rather than cooperative? In a repeated coordination game, pairs received advice that was either cooperative or selfish with common ground manipulated by whether pairs knew they received identical advice. We found that advice type shaped the pattern of coordination: pairs with selfish advice converged toward asymmetric equilibria, while pairs with cooperative advice maintained symmetry. Common ground amplified this pattern: pairs who knew they shared selfish advice showed the steepest increase in round-level payoff asymmetry as rounds progressed. We interpret this as recursive reasoning shaping equilibrium selection: common ground on selfish advice, increases expectations of selfishness, making acceptance of asymmetry rational—anticipation reinforcing the pattern. Linguistic analysis complements this, where participants' use of "we" tracked fairness outcomes. These findings demonstrate that common ground shapes not just whether coordination succeeds, but the kind of coordination.

Keywords: Common Ground; Coordination; Information Transmission;

Introduction

Successful coordination depends on shared expectations. When I am able to anticipate your behavior and you are able to anticipate mine, we can align our actions and converge on mutually beneficial outcomes (Clark, 1996; Lewis, 1969). For example, the recursive structure of common knowledge (Aumann, 1976; Geanakoplos, 1992) is what allows strangers to meet at Grand Central Station without needing to explicitly communicate about it (Schelling, 1960). Experimental work has confirmed that people are sensitive to this recursive structure: for instance, participants are more likely to attempt risky coordination when payoffs are broadcast publicly than when they are merely shared (i.e. each person knows privately; Thomas et al., 2014). Common knowledge thus appears to function as a distinct cognitive state that humans deploy to solve coordination problems (De Freitas et al., 2019; DeScioli et al., 2014).

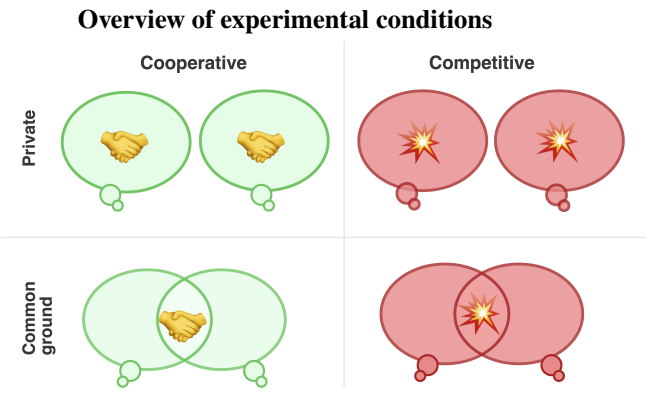
But does common ground help coordination regardless of what is shared, or does it depend on the *content* of the shared expectations? Most demonstrations of common ground involve neutral or cooperative focal points—a time, a place, an arbitrary label (Mehta et al., 1994; Schelling, 1960). What happens when the shared information is competitive rather than cooperative? Consider advice to "get the best outcome for yourself" versus "find a fair outcome for both of you." Three views make different predictions.

One possibility is that common ground helps regardless of content. Even selfish advice provides a shared reference point that reduces strategic uncertainty, allowing players to coordinate on some equilibrium rather than none.(cf. Chaudhuri et al., 2009). A second possibility is that the content of advice matters, but common ground does not interact with it: competitive framing may prime selfish behavior through individual-level mechanisms like construal or goal activation (Liberman et al., 2004), but this effect operates independently of whether the framing is shared. A third possibility is that common ground shapes whatever content is shared, by steering pairs toward different equilibria. When I know you received selfish advice, I may expect you to pursue high individual payoffs; knowing that I know, you expect the same of me (cf. Fischbacher et al., 2001). We call this *destructive recursive reasoning*: a self-fulfilling prophecy in which common knowledge of competitive framing lead us to accept unequal outcomes that neither would tolerate if the competitive framing were private.

Computational models of social learning typically treat cooperativeness as something inferred from a partner's behavior over time. The Bayesian Reciprocator, for example, uses theory of mind to infer latent preferences from observed actions, adjusting cooperation as evidence accumulates (Kleiman-Weiner et al., 2025). But expectations often precede interaction entirely. Pre-play information can shape initial beliefs before any behavioral evidence is available, whether it comes in the form of a situational label (Liberman et al., 2004), selective reporting about others' past behavior (Engel et al., 2021; Snyder & Haugen, 1995), or explicit advice (as we provide). If these initial beliefs drive initial behavior, and initial behavior shapes subsequent dynamics, then the source of expectations may matter as much as the expectations themselves. We ask what happens when competitive expectations are not only present but mutually known: when I know you received selfish advice, and you know I know, recursive reasoning may lock us into accepting inequality before we ever have a chance to negotiate otherwise. When a public event gives each player reason to believe the other will act selfishly, and each knows the other has this reason, the recursive chain of "I believe you believe I believe..." may unfold automatically (Cubitt & Sugden, 2003).

We test this possibility in a repeated coordination game

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



Instructions by condition

Condition	Advice
Selfish	"Try to get the best outcome for yourself."
Cooperative	"Try to find an outcome that's fair for both of you."

Condition	Note Displayed
CG	"Each pair gets the same advice, so your partner will see exactly the same advice as you do."
Non-CG	"The advice is assigned randomly so your partner may have received advice from someone else."

Figure 1: *2 by 2 experimental design.* The left figure illustrates the 2 by 2 design where advice was either cooperative or competitive, and either private (each player's own knowledge) or common ground (both players knew they shared the same advice). The right tables show the exact phrasing of instructions in each condition.

adapted from Yu and Thompson, 2024 where pairs of participants choose between options that vary in how they balance individual versus joint payoffs. Before playing, participants receive advice emphasizing either cooperation or competition. Crucially, we manipulate whether this advice is common ground: some participants know their partner received identical advice, while others do not know what advice their partner received. In a series of exploratory analyses, we found that the interaction of common ground and advice type lead to significant differences in coordination dynamics. Participants who received selfish advice chose high-payoff options more frequently, consistent with a utility model in which cooperative agents avoid actions that guarantee they receive more than their partner. While all conditions successfully avoided collisions at similar rates, pairs diverged in how they coordinated: those with selfish advice converged toward asymmetric equilibria, while those with cooperative advice maintained symmetry. This divergence was most pronounced when advice was common ground: pairs who knew they shared selfish advice showed the steepest increase in payoff asymmetry over rounds, while pairs who knew they shared cooperative advice maintained symmetry throughout. This suggests common ground shapes not *whether* coordination succeeds, but what *kind* of coordination emerges, pointing to interesting future work as well as modeling implication.

Experiment

Following Yu and Thompson, 2024 we conduct a repeated coordination game in which participants coordinate where to park their cars.

Methods

Design and Materials We used a 2 (advice type: cooperative vs. selfish) × 2 (common ground: CG vs. noCG) between-subjects design. Pairs played 13 rounds of a coordination game in which each player simultaneously chose one of four parking spots (A, B, C, or D). Payoffs depended on both players' choices, which were explained to the participants through both text and illustrated examples (Fig. 2). If both chose the same spot, they collided and earned nothing. If they chose different spots of the same color, they received a coordination bonus. The game has two Nash equilibria, corresponding to the highlighted cells in Table 1 (A Nash equilibrium in this setting is a pair of actions choices made by the dyad such that for each participant, the best action that they can take given their partner's action is their current action): the *symmetric* equilibrium (A, B), where both earn 20 points, and the *asymmetric* equilibrium (C, D), where one earns 29 and the other 17. This creates tension between efficiency and fairness. The asymmetric equilibrium has a higher joint payoff, but the symmetric equilibrium avoids one being better off than the other. Before playing, participants received advice ostensibly from a previous participant: either "Try to find an outcome that's fair for both of you" (cooperative condition) or "Try to get the best outcome for yourself" (selfish condition). We manipulated common ground by telling participants either that their partner received identical advice (CG) or that advice was randomly assigned (noCG).

Participants We recruited 445 participants across all conditions via prolific. We lost some participants due to disconnections, as well as failing attention checks. The final number of pairs per condition were n=35,35,36,36 for CG_cooperate, CG_selfish, noCG_cooperate, and noCG_selfish respectively. Participants were paid $2 base rate plus a bonus payment depending on their performance, their median completion time

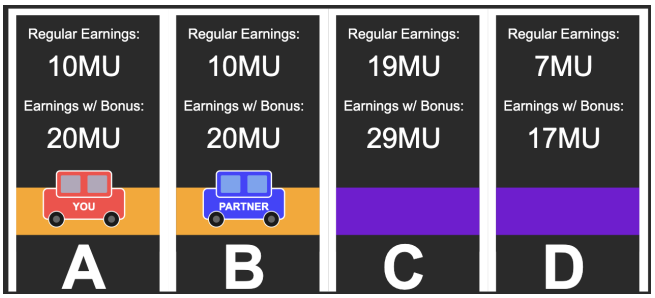


Figure 2: *Instruction illustration*. Participants were shown example payoffs.

		Player 2			
		A	B	C	D
Player 1	A	(0,0)	(20,20)	(10,19)	(10,7)
	B	(20,20)	(0,0)	(10,19)	(10,7)
	C	(19,10)	(19,10)	(0,0)	(29,17)
	D	(7,10)	(7,10)	(17,29)	(0,0)

Table 1: *Payoff matrix*. Diagonal cells (gray) are collisions yielding zero. Nash equilibria are highlighted.

was <10min, and participants received bonus payments depending on how many MU they collected in the task. Specifically, each MU corresponded to $0.005

Procedure Participants were paired online, read instructions, received the advice manipulation, and wrote a brief game plan. They then completed 13 rounds of the game (90 seconds per round), with feedback after each round of their payoff. Attention checks every 4 rounds asked participants to indicate the bonus they earned in the previous round. Following the game, participants in the CG condition answered: "Do you feel you and your partner followed the advice well?"

Predictions We expect agents to be informing their decisions based on relative rewards between themselves, R_{self}, and their partner, $R_{partner}$. We make two key predictions of how this decision might depend on their cooperativeness: (1) Selfish agents will choose C more, because choosing C guarantees dominance over their partner (i.e. $R_{self} \geq R_{partner}$) regardless of their partner's action, creating a fairness penalty only for cooperative agents. Cooperative agents will not want to appear selfish to their partner, and hence will not choose C. (2) Selfish agents will tolerate payoff differences (i.e. coordinating on purple), while cooperative agents should prefer the equal equilibria.

Results

Terminology We clarify the terminology used to discuss our results as they are easily conflated. We use "selfish" and "cooperative" to refer to the two types of advice given to participants, "symmetric" and "asymmetric" to refer to the type of equilibrium (corresponding to the orange and purple shaded Nash equilibria respectively in Table 1) and equal and unequal to refer to the difference in total bonus outcome between the two players in the same dyad.

Behavioral Analyses

First, we confirm previous findings (Yu & Thompson, 2024) that as rounds progressed, participant's bonuses improved ($p < 0.001$). This suggests that participants coordinated on strategies in earlier rounds, which then stabilized over time to successfully coordinate.

Bonus differences were similar between conditions. To see how different conditions affect coordination success, we fit a linear model with common ground and advice type as fixed effects, we found that common ground significantly improved coordination across the board (β=19.12, t=2.00 p=.047) with participants in the CG condition achieving an average of 193MU per game while participants in the non-CG condition achieved 174MU. However, using the same model to predict bonus differences between pairs, we found no significant effect of common ground nor advice type on overall bonus difference between participants, contrary to preliminary predictions.

Participants receiving selfish advice chose C more. We investigate our prediction of difference in tendencies to select C. Using a generalized mixed-effects linear model with common ground, advice type and round as fixed effects with random intercept for each player, we find that participants who received the selfish advice chose C more($\beta = 0.53, z = 2.11, p = .035$). Despite this, there is no significant difference in the rate of collision across conditions, suggesting that beyond avoiding collisions, there was motivation not to park in C in the cooperative condition supporting our model.

Asymmetries in equilibria selection depended on an interaction of CG, advice type, and round Conditions having no significant difference in overall bonus inequalities suggests that advice and common ground do not change people's underlying tendencies to be "fair". However, difference in the frequencies of choosing action C indicates that action patterns differ between conditions. We investigate this more closely by looking at round-level payoff differences - a crucial differentiating factor between the two possible equilibria, as explained in the Design and Materials section.

Fitting a linear mixed-effects model predicting round-level payoff differences from advice type, common ground, round, and their interactions, with random intercepts for each pair, we find a 3-way interaction between advice type, common ground and round (β = -1.12, t = -2.17, p = 0.032; Figure 3). Analyzing the slope of each condition using 'emtrends' we find that the coefficient for the CG+ Selfish condition was estimated to be the highest (β = 0.255 , 95%CI = [0.118,0.392]) while CG + Cooperative was the lowest, with a near zero difference across rounds (β=0.016, 95%CI=[-.122,.153]) suggesting that common ground made the effects of advice more pronounced in shaping the equilibrium that pairs converge to. The fact that round-level differences are similar at the start when players are negotiating coordination strategies, but diverge overtime indicate that the advice has a lasting effect overtime which only become apparent once strategies stabilize.

Strategy Classification

Pairs in the CG + Selfish condition mitigated unfairness by alternating. The analyses so far indicate that the CG + Selfish condition converges to the asymmetric payoff equilibrium, and yet this does not lead to any significant difference in overall bonuses inequalities between conditions. This suggests that

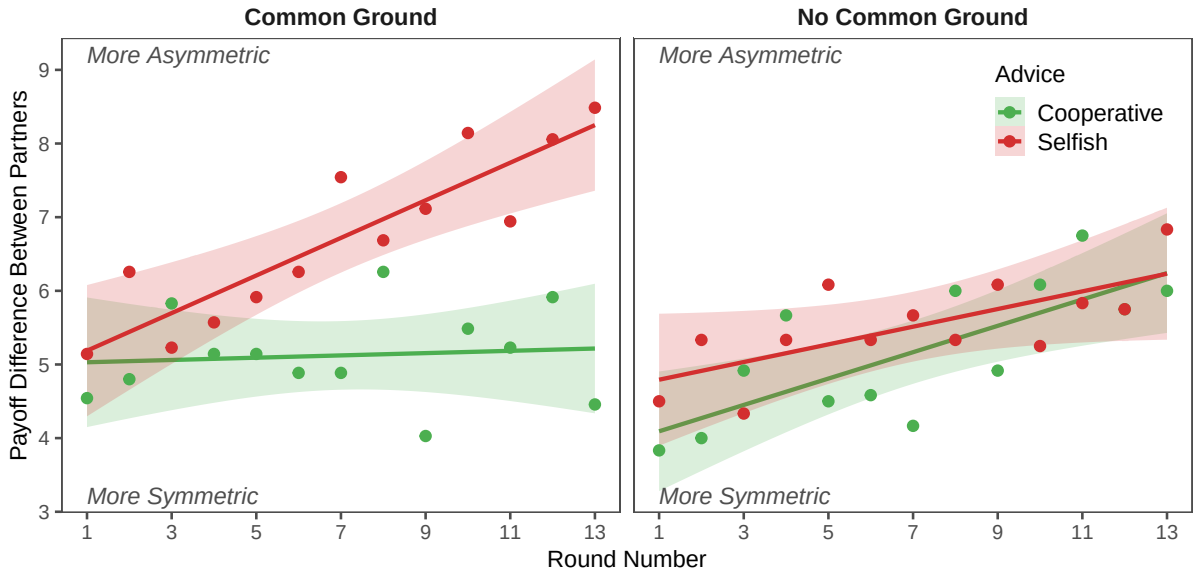


Figure 3: *Payoff asymmetries*. Points are round-level means; lines are linear fits with 95% CI ribbons. The CG + Selfish condition shows increasing round-level asymmetry while CG + Cooperative remains flat.

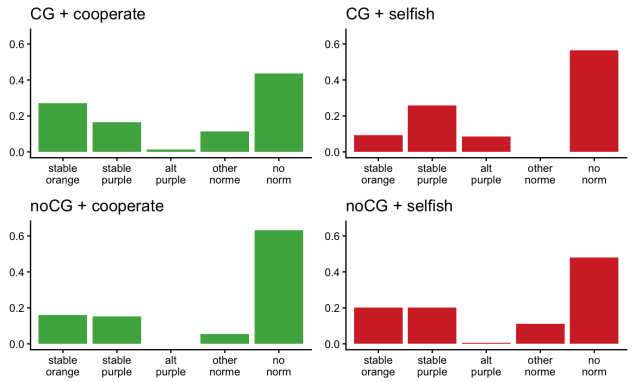


Figure 4: *Strategy frequencies by condition.* Common ground increased convergence to stable strategies, with cooperative advice favoring the symmetric equilibrium (stable orange).

participants in the CG + Selfish condition may be compensating for the asymmetric payoffs in another way, namely by alternating. We thus investigate the strategies participants are adopting more closely, by categorizing them into "stable orange", "stable purple" and "alternating purple", "other norm" and "no norm" based on the last 4 rounds as done in Yu and Thompson, 2024(see figure 4). Using a general linear model predicting alternation of C and D from advice type, common ground, and their interactions, we find as predicted that the CG + Selfish condition had significantly higher frequencies of alternating purple, than in the CG + Cooperative condition ($\beta = 1.87, z = 2.41, p = 0.02$). This explains how despite round level payoffs increasing in the CG + Selfish condition as seen in figure 3, they have similar total bonuses by alternating on the higher payoff.

Advice shapes equilibrium selection, and common ground enhances it. We further investigate our hypothesis that advice shapes which equilibrium participants coordinate on - the symmetric (stable orange) or the asymmetric (alternating purple or stable purple). Using the same general linear model to predict the occurrences of the stable orange strategy, we find significance for both fixed effects on advice type ($\beta = 1.29, z = 3.72, p < 0.001$) and common ground ($\beta = 0.673, z = 2.27, p = 0.02$), as well as their interaction ($\beta = 1.57, z = 3.39, p < 0.001$). Pairwise comparisons using emmeans revealed that pairs in the CG + Cooperative condition were significantly more likely ($\beta = 1.29, z = 3.72, p = 0.001$) than CG + selfish to exhibit stable orange while noCG + Cooperative and noCG + Selfish were not significantly different. A similar analysis looking at the frequency of asymmetric payoff showed that selfish advice increased chances of pairs converging on the asymmetric equilibria ($\beta = 0.88, z = 3.09, p = 0.002$). Pairwise comparisons revealed that CG + Cooperative was significantly less likely to converge to an asymmetric equilibria than CG + Selfish($\beta = -0.875, z = -3.088, p = 0.01$), while the conditions, noCG + Cooperative and noCG + Selfish were not significantly different, suggesting that common ground increased likelihood of converging to a norm overall, driven by the cooperative condition, and thus exhibit difference in norm convergence, whereas the noCG conditions didn't converge to a norm so didn't exhibit these differences. Investigating how often by making the target of the model the 'no norm' category further confirms this; We found significant fixed effects ($\beta = 0.517, z = 2.145, p = 0.03$ for the advice type) and interaction ($\beta = 1.14, z = 3.35, p < 0.001$), and most relevantly, revealed that having common ground significantly increased the likelihood to converge to a norm

($\beta = 0.80, z = 3.30, p < 0.001$).

Overall, strategy categorization confirms that advice changed the kinds of equilibrium dyads converged to. This difference was enhanced by common ground suggesting that this allowed dyads to coordinate and converge to strategies more reliably.

Linguistic Analyses

The behavioral analyses show that common ground amplifies the effect of advice on equilibrium selection, but they do not reveal *why*. Two accounts are consistent with these results. First, participants may infer their partner's likely behavior from the advice and update their own strategy accordingly (e.g. "I'm going to be selfish because they will be"). This nudges participants to choose some actions over others (e.g. choosing the purple parking slots instead of orange) which by the last rounds of the experiment leads to convergence on a particular equilibrium (e.g. asymmetric). A second, more impartial account of their behavior is that participants use shared advice primarily as a coordination device, a focal point for agreeing on an equilibrium. Especially for the CG_selfish condition, it is obvious that both participants choosing the highest payoff parking lot will be unproductive. Thus, they may use the advice to help coordinate their actions (e.g. choose the purple, asymmetric equilibrium and alternating) without genuinely taking the advice to heart.

To distinguish these accounts, we analyzed participants' written responses before and after the game, focusing on pronoun usage as a window into how participants framed the interaction. Prior work suggests that pronoun use reflects self-focus (Kacewicz et al., 2014), and individualistic vs collectivist orientation (Na & Choi, 2009), while recent work on coordination suggests that successful joint action involves creating a shared "we" agent (Tang et al., 2025). We thus counted uses of "I" and "we" in participants' responses to examine whether they framed the game as joint coordination or individual action. As we did not have a survey prompt for the noCG condition, these analyses are restricted to the common ground condition.

Initial framing reflects advice. In responses to the pre-game prompt "What's your game plan?" participants in the cooperative condition used "we" more frequently than those in the selfish condition ($\beta = 1.66, z = 3.04, p = .002$). This suggests that advice shaped participants' initial framing of their partner. However, this difference is difficult to interpret cleanly: the cooperative advice explicitly mentioned "both you and your partner," while the selfish advice did not, potentially priming partner-oriented language independent of participants' actual intentions.

Final framing reflects outcomes. More informative is what happened after the game. In response to the exit prompt "Do you feel you and your partner followed the advice well?" advice type no longer predicted "we" usage. Instead, we found an effect of outcome inequality. Participants who achieved more equal overall outcomes used "we" more frequently ($\beta = -0.005, z = -2.13, p = 0.033$; see Figure 5). Conversely, initial "we" usage did not predict eventual fairness outcomes ($p = 0.144$). This pattern is more consistent with the reasoning-about-partner account than with advice-as-focal-point. If participants were simply using advice as a coordination device, we would expect initial framing to persist regardless of game dynamics. Instead, participants' sense of shared intentionality tracked what actually happened during play, suggesting they were attending to their partner's behavior and updating their interpretation of the interaction accordingly. That payoff *inequality*, rather than absolute payoffs, predicted "we" usage further suggests that participants spontaneously tracked fairness even when not explicitly asked about it.

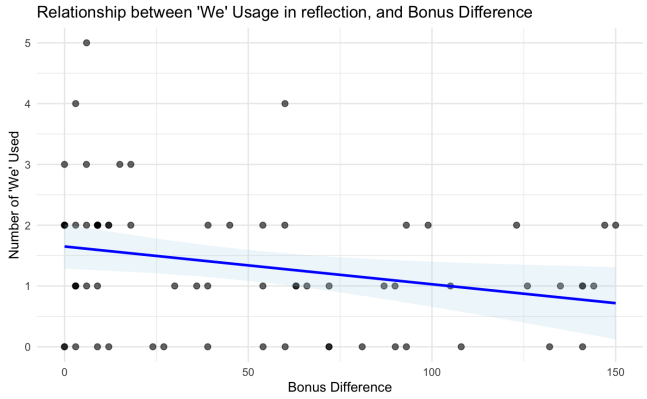


Figure 5: Number of uses of "we" in exit survey against total bonus difference

General Discussion

Common ground is typically studied as a solution to a binary problem: will coordination succeed or fail? But coordination games often have multiple equilibria. Our findings suggest that common ground shapes equilibrium selection by making shared expectations self-fulfilling. When participants knew they shared cooperative advice, they maintained equality; when they knew they shared competitive advice, they drifted toward asymmetry. Common ground helped coordination in both cases, as pairs with shared knowledge of any kind outperformed those without, but it helped them coordinate on different things. This distinction re-frames what common ground does. Prior work has emphasized that common knowledge enables coordination that would otherwise fail: people attempt risky coordination when payoffs are public but not when they are merely shared (DeScioli et al., 2014; Thomas et al., 2014). Here, pairs coordinated successfully regardless of condition. Collision rates were similar throughout. What differed was the equilibrium they converged on. Common ground functioned less as a switch that turns coordination on, and more as a steering mechanism that directs it.

Why did advice only matter when it was common ground?

If competitive framing worked through individual-level mechanisms (goal activation, construal, or priming) we would expect effects regardless of what participants knew about their partners. The absence of differentiation in the noCG conditions suggests that the leverage common ground provides to predict how the partner would reason induces a mechanism of recursive theory of mind reasoning between the dyad: I adjust my behavior because I expect you to adjust yours, and I expect you to adjust because I know you know we received the same advice. Without that mutual knowledge, the advice is just information about the game; it does not license inferences about what my partner will do, so it provides no basis for strategic adjustment. This is consistent with formal accounts of common knowledge as a distinct epistemic state that supports coordinated action in ways that mere shared information cannot (Cubitt & Sugden, 2003).

Models of cooperation and coordination These dynamics have implications for how we model the emergence of cooperation. Computational approaches typically treat a partner's cooperativeness as something inferred from observed behavior over time (Kleiman-Weiner et al., 2025). But if preplay information shapes initial behavior, and initial behavior shapes what partners infer about each other, then early trajectories may be hard to escape. A pair that starts with mutual expectations of selfishness may behave selfishly, confirm each other's expectations, and continue, even if both would have cooperated under different initial conditions. This suggests that where cooperation comes from may depend not just on incentives or learning, but on what is mutually known at the outset. More generally, this work provides insight into how people balance competition and cooperation - a tension which is present in many real world settings. Economic(Fischbacher & Gachter, 2008) and reinforcement learning models of cooperation often ascribe this complexity to the ability of agents to recursively reason about other agents in order to maximize their own reward. However, other models ascribe cooperativeness to their construal of their partner and themselves as a joint agent (McKee et al., 2023; Yu and Thompson, 2024) intrinsically motivated to maximize their joint reward. Our results suggest that both are important. Common ground's effect on coordination dynamics suggest that how participants recursively reason about their partner is important in explaining their coordination behavior. On the other hand, our analysis of "we" suggest how participants' perceptions may be affected by fairness which may in turn affect their coordination dynamics. Future work could explore this interaction, of how people's social utilities (to be fair, respectable etc.) and their drive to maximize reward interact to form the coordination dynamics we observe, where people are reasoning about themselves jointly. Capturing this dynamic will help build models which cooperate better with humans, as noted in Kleiman-Weiner et al., 2025

Iterated Learning Our results show that the way advice is presented influences the outcome, indicating implications for models of iterated learning and evolution. Typically models of information transmission is modeled in terms of the amount of information (Imel et al., 2023; Kirby et al., 2015), but our work suggests that advice is more than its information content. When advice is shared publicly, it may become a coordination device, set initial expectations or start self-fulfilling dynamics. While there is some work on how the way information is transmitted affects learning (Shafto et al., 2014; Thompson et al., 2022), researchers should be mindful of how nuances or transmission, such as creation of common ground through the act of giving advice, affects the coordination dynamics of the next generation. Future work should incorporate the effect of the presentation of advice in influencing coordination in models iterated learning in a population.

Limitations and next steps Firstly, some participants struggled with the attention check, suggesting that the rules and payoffs were complicating. This affected all manipulation types equally but ensuring participants understood the task through more test examples may have led to cleaner results. The cooperative advice explicitly mentioned "both you and your partner," while the selfish advice did not, conflating competitive framing with reduced partner-salience. Future work should cross these factors independently. Other than increased information about the partner leading to more recursive reasoning, there are multiple possible mechanisms of how participants may be reasoning about common ground. For example, common ground may be creating an atmosphere of cooperation or selfishness without triggering recursive reasoning. Conducting tighter experiments to distinguish this account, such as comparing cooperative dynamics between the case where the advice is in common ground with the case where you know your partner's advice, and also that your partner doesn't know your advice, will help us better understand which mechanism the observed effect in this experiment. We partially tried to address how individuals were reasoning about their partner through linguistic analyses of the use of "we", but more direct measures of belief attribution would strengthen this interpretation. Furthermore, to build a more general understanding of the mechanism of cooperative dynamics and make progress towards building a computational model, future work could extend beyond this particular paradigm to include different environments which differ in aspects such as payoffs, types of communication, and the methods of information transmission and learning, in order to investigate their effect on influencing behavior. Ultimately, it remains to be investigated whether the same dynamics arise when competitive expectations become common knowledge through subtler channels, from organizational culture or media framing to historical precedent. But if they do, such expectations may be self-fulfilling in ways that are difficult to reverse.

All code and materials available at:
https://github.com/yukam997/
CommonGroundShapesCoordination

Acknowledgments

Yuka Machino was supported by the Ezoe Memorial Recruit Foundation, as well as the HUMAI program[1]. We thank the anonymous reviewers for their feedback. AI was used to help for coding tasks including generating parking images for the study, generating graphs, data wrangling and to brainstorm background literature.

References

Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics*, *4*(6), 1236–1239.

Chaudhuri, A., Schotter, A., & Sopher, B. (2009). Talking ourselves to efficiency: Coordination in inter-generational minimum effort games with private, almost common and common knowledge of advice. *The Economic Journal*, *119*(534), 91–122.

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

Cubitt, R. P., & Sugden, R. (2003). Common knowledge, salience and convention: A reconstruction of david lewis' game theory. *Economics and Philosophy*, *19*(2), 175–210.

De Freitas, J., Thomas, K., DeScioli, P., & Pinker, S. (2019). Common knowledge, coordination, and strategic mentalizing in human social life. *Proceedings of the National Academy of Sciences*, *116*(28), 13751–13758.

DeScioli, P., Haque, O., & Pinker, S. (2014). The psychology of coordination and common knowledge. *Journal of Personality and Social Psychology*, *107*, 657–676. https://doi.org/10.1037/a0037037

Engel, C., Kube, S., & Kurschilgen, M. (2021). Managing expectations: How selective information affects cooperation and punishment in social dilemma games. *Journal of Economic Behavior & Organization*, *187*, 111–136.

Fischbacher, U., & Gachter, S. (2008). Social preferences, beliefs, and the dynamics of free riding in public Good experiments. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1314687

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics letters*, *71*(3), 397–404.

Geanakoplos, J. (1992). Common knowledge. *Journal of Economic Perspectives*, *6*(4), 53–82.

Imel, N., Futrell, R., Franke, M., & Zaslavsky, N. (2023). Noisy population dynamics lead to efficiently compressed semantic systems. *NeurIPS 2023 workshop: Information-Theoretic Principles in Cognitive Systems*.

[1]https://zen.ac.jp/humai/project1

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2014). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, *33*(2), 125–143. https://doi.org/10.1177/0261927X13502654

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102. https://doi.org/10.1016/j.cognition.2015.03.016

Kleiman-Weiner, M., Vientós, A., Rand, D. G., & Tenenbaum, J. B. (2025). Evolving general cooperation with a Bayesian theory of mind. *Proceedings of the National Academy of Sciences*, *122*(25), e2400993122. https://doi.org/10.1073/pnas.2400993122

Lewis, D. (1969). *Convention: A philosophical study*. John Wiley & Sons.

Liberman, V., Samuels, S. M., & Ross, L. (2004). The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. *Personality and social psychology bulletin*, *30*(9), 1175–1185.

McKee, K. R., Hughes, E., Zhu, T. O., Chadwick, M. J., Koster, R., Castaneda, A. G., Beattie, C., Graepel, T., Botvinick, M., & Leibo, J. Z. (2023). A multi-agent reinforcement learning model of reputation and cooperation in human groups. https://arxiv.org/abs/2103.04982

Mehta, J., Starmer, C., & Sugden, R. (1994). The nature of salience: An experimental investigation of pure coordination games. *The American Economic Review*, *84*(3), 658–673.

Na, J., & Choi, I. (2009). Culture and first-person pronouns [PMID: 19713570]. *Personality and Social Psychology Bulletin*, *35*(11), 1492–1499. https://doi.org/10.1177/0146167209343810

Schelling, T. C. (1960). *The strategy of conflict*. Harvard University Press.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89. https://doi.org/10.1016/j.cogpsych.2013.12.004

Snyder, M., & Haugen, J. A. (1995). Why does behavioral confirmation occur? A functional perspective on the role of the target. *Personality and Social Psychology Bulletin*, *21*(9), 963–974. https://doi.org/10.1177/0146167295219009

Tang, N., Gong, S., Zhao, M., Zhou, J., Shen, M., & Gao, T. (2025). Joint commitment in human cooperative hunting through an 'imagined we'. *Proceedings. Biological sciences*, *292 2055*, 20243070. https://api.semanticscholar.org/CorpusID:281504623

Thomas, K. A., DeScioli, P., Haque, O. S., & Pinker, S. (2014). The psychology of coordination and common knowledge. *Journal of personality and social psychology*, *107*(4), 657.

Thompson, B., van Opheusden, B., Sumers, T., & Griffiths, T. L. (2022). Complex cognitive algorithms preserved by selective social learning in experimental populations. *Sci-*

ence, *376*(6588), 95–98. https://doi.org/10.1126/science.abn0915

Yu, D., & Thompson, B. (2024, June). People balance joint reward, fairness and complexity to develop social norms in a two-player game. https://doi.org/10.31234/osf.io/qe6bp