Title
Let's be friends! People work together even when there is no incentive to do so

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors
Btesh, Victor
Ho, Mark

Publication Date
2026

Let's be friends! People work together even when there is no incentive to do so

Victor Btesh[1*], **Mark K. Ho**[2*]
[1]Department of Experimental Psychology, University College London
[2]Department of Psychology, New York University
*vbtesh9@gmail.com

Abstract

One distinguishing feature of human social intelligence is our capacity to flexibly coordinate complex behaviors, such as when working together to prepare a meal. What cognitive or motivational processes make such joint action possible? Here, we investigate the role that a *preference to enmesh behaviors* plays in human coordination using a novel two-player virtual baking paradigm. Participants could either independently bake bread by bringing their own water and flour, or jointly bake bread by bringing only water or only flour to share. Depending on the ingredient options available, trials in our experiment could correspond to either the classic Prisoner's Dilemma, in which acting jointly is disincentivized by the payoffs, or a "Friend's Dilemma", in which acting jointly is neither incentivized nor disincentivized (though it remains risky since one's partner might not share). Formal analysis of equilibria and dyadic learning dynamics indicate that people work together even when there is no incentive to do so. These results provide preliminary evidence that a motivation to enmesh behaviors with social partners supports successful coordination.

Keywords: relationships; social cognition; cooperation; game theory

Introduction

Consider baking bread with a friend. You could each bake your own loaves, same flour, same recipe, same outcome. Nothing particular is gained by combining your efforts. If anything, coordinating adds the risk that one of you fumbles and wastes the other's ingredients. Yet, when a friend suggests baking together, it likely feels more appealing that doing so alone. We do not experience this as irrational. It feels obviously, unambiguously better; not because of the bread, but because of the company. This intuition sits uneasily with formal theories of human behaviour. From a utility-maximising perspective, a rational agent should be indifferent between baking alone and baking together when the expected payoffs are identical. And yet we are not. We choose to enmesh.

This intuition has empirical support. Tomasello (2008) proposes that from childhood, humans hold not only the ability for joint intentionality but an intrinsic preference for collaborative action. This has been shown to be a key differentiating feature between humans and other large apes (Warneken & Tomasello, 2006; Warneken et al., 2006). It suggests that, from a young age, humans have can represent goals jointly with others and seek cooperation even when others are not needed to achieve goals (Warneken et al., 2012). Humans cooperate not for gains but for its own sake. Historically, these accounts have lacked formal definitions, and have consequently been ignored in game theoretic models. Still, several computational frameworks explain cooperation despite incentives to defect, including models of indirect reciprocity sustained through reputation (Hilbe et al., 2018; Nowak & Sigmund, 2005; Santos et al., 2021; Schnell & Muthukrishna, 2024). Kleiman-Weiner et al. (2025) argue that theory of mind could be the cognitive engine over which indirect reciprocity operates. Yet none address how the drive to enmesh is represented in everyday experience: what makes baking together *feel* more rewarding than doing it alone?

Classical accounts, where cooperation stems from calculated expected gains (Becker, 1976; Von Neumann & Morgenstern, 1944) would consider irrational a preference for doing things together. Alternatively, virtual bargaining proposes that players simulate the outcome of explicit negotiation (Chater et al., 2022; Le Pargneux et al., 2024), but does not that players merge their utilities. Closer to the target, team reasoning proposes that players build a joint utility function to select strategies (Colman, 2003; Colman et al., 2008). Both accounts explain cooperation in games which discourage cooperation through risk such as the Prisoners' dilemma and Stag Hunt (Colman & Gold, 2018; Rapoport & Chammah, 1965; Skyrms, 2003). While team reasoning explains how agents reason once identified as a team, it takes team identification as exogenous rather than a choice. It stays silent about whether people prefer to join a team in the first place. The claim advanced by Tomasello (2008) is more radical: humans hold an intrinsic preference for cooperation and joint action, a drive to form teams that precedes any strategic calculation about the benefits of doing so.

This preference makes a testable prediction: people should choose to enmesh even when doing so provides no material advantage or is risky. We do not merely tolerate interdependence; we seek it. In this paper, we introduce the Friends' dilemma, a game designed to isolate this preference by eliminating any material advantage to cooperation. We find that participants systematically choose to enmesh, taking unrewarding risks. Furthermore, this choice is necessarily associated with feeling closer to one's partner, a link absent in games where cooperation is payoff rational.

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

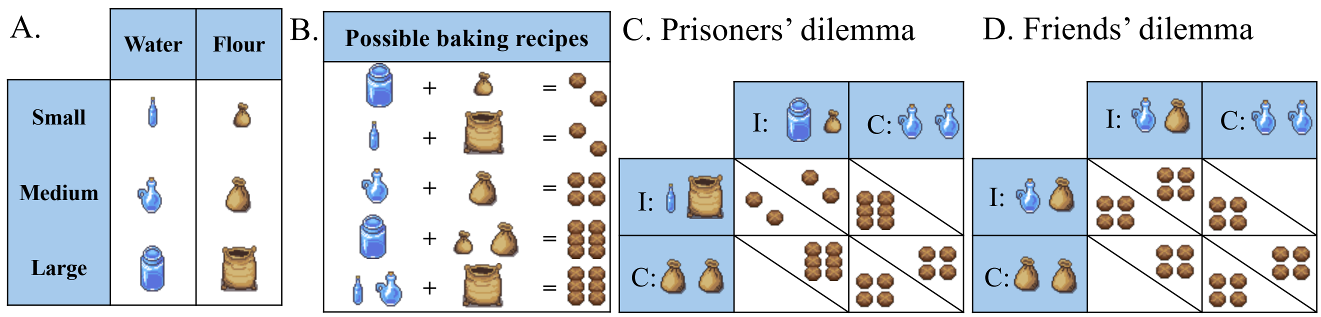

Figure 1: **Specifications of the virtual baking paradigm.** A. Possible quantities of flour and water in the baking game. B. Possible number of breads resulting from combining ingredients of different sizes. The ingredient which has the smallest quantity determines the outcome. The list does not show all recipes but one way to reach each possible outcome. C. Prisoners' dilemma in the baking game, ignoring the second step where participants choose to give, take or stay idle for clarity. D. Friends' dilemma in the baking game, where the payoffs for acting independently and collaborating are equal.

Enmeshment in the Friends' Dilemma

In this work, we examine people's behavior in the *Friends' Dilemma*, a two-player game designed to evaluate people's preference for enmeshment. The game is embedded in a baking scenario designed to make the strategic structure intuitive to participants.

Virtual Baking Paradigm Each player begins with ingredients, flour and water, that can be combined to make bread. Crucially, bread production is limited by whichever ingredient is smallest: having more flour yields nothing without the water to match it (Figure 1 A, B). The game proceeds in two steps. In the first step, players simultaneously choose to be *independent* (I) or *cooperate* (C), yielding strategies $S_1 = \{I, C\}$. Being independent means bringing ingredients of each type: one water and one flower, and baking independently from the other player. Collaborating means bringing two ingredients of a single type and pooling them with the other player. If both cooperate, each player gets an equivalent quantity of water and water; if only one cooperates, their contribution is wasted and they cannot bake. The second step, which we call *give and take*, allows players to redistribute the bread produced in step one. Players simultaneously choose to *take* (T) one bread from the other, remain *passive* (P) or *give* (G) one of their own, yielding strategies $S_2 = \{G, P, T\}$ (Figure 2). The full strategy space is thus $S = S1 \times S2 = \{IT, IP, IG, CT, CP, CG\}$.

The baking game can be configured as either a Prisoners' dilemma (PD) or a Friends' dilemma (FD) depending on the ingredient quantities available (Figure 1). When pooling ingredients yields more bread than keeping them separate, cooperation is risky but rewarding: this is PD, and frameworks such as team reasoning and virtual bargaining correctly predict cooperation (Chater et al., 2022; Colman & Gold, 2018). When pooling and keeping yield equal bread, cooperation is risky yet unrewarding: this is FD. Both games share a Nash equilibrium when both players choose to be independent (followed by take), which risk dominates other strategies

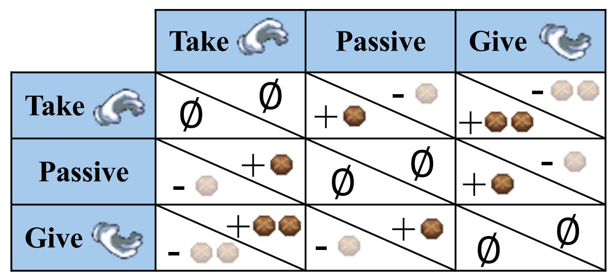

Figure 2: **Payoffs matrix for the second step of virtual baking paradigm**. Note that the plus and minus sign represent the fact that these payoffs are conditioned on the number of breads players have baked in the first round. If both players have 4 breads, then this matrix applies as is. If a player has 0 bread and the other has 4 breads, then the player with 0 bread can try to give, but this will have no effect unless the other baker gives as well: you can give what you do not have. Similarly, the player with 4 cannot take from the player with 0 breads. Choosing symmetric actions always yields unchanged payoffs as take-take and give-give cancel out.

(Harsanyi & Selten, 1988). The games diverge, however, in what other theories predict. In PD, mutual separation is not Pareto optimal, so virtual bargaining predicts a jump to cooperation. In FD, mutual separation already lies on the Pareto frontier: the disagreement point offers the same payoff as any agreement. Virtual bargaining is therefore indifferent among all symmetric outcomes, and team reasoning faces the same indifference since joint payoffs are identical across equilibria. Neither theory provides a reason to cooperate, nor to favour taking, giving, or staying passive in the second step.

Theoretical predictions This indifference is precisely what makes the friends' dilemma diagnostic. Nash makes a clear predictions that people will be independent in both games. Because team reasoning and virtual bargaining are neutral between all six equilibria in FD, any systematic preference for one over the other must stem from something else (see

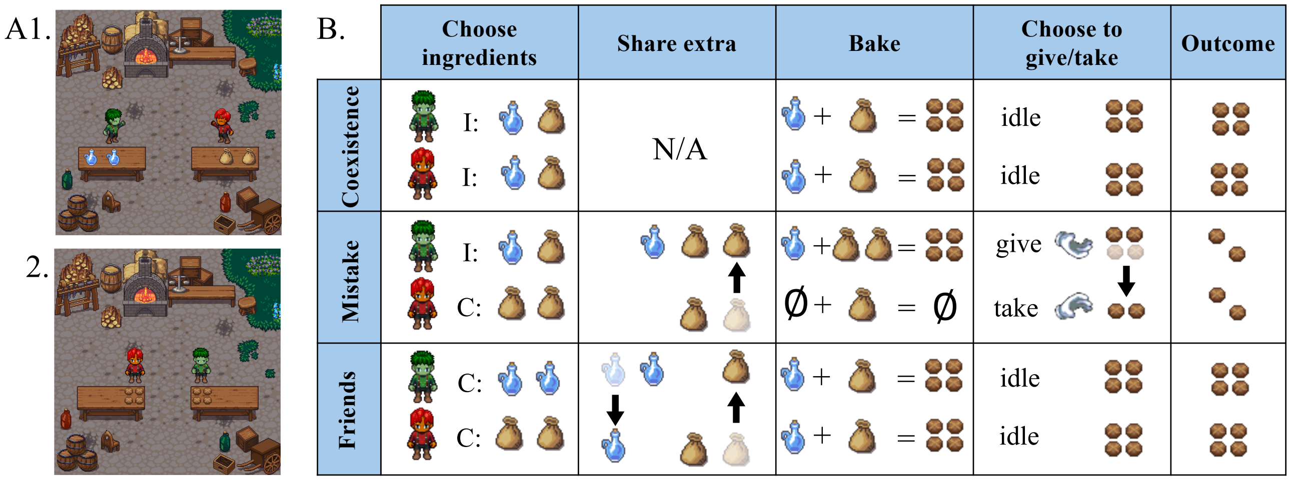

Figure 3: **Materials and schematic of each possible outcome in the Friends' dilemma** A. Screenshots from the videos depicting peoples actions as participants saw them. Here, two players are cooperating and get each 4 breads. B. Example rounds of the Friends' dilemma showing the possible states. Notice how coexistence and friends offer the same payoffs and how participants can use the give/take step to correct mistakes and rebalance outcomes should they wish to.

Table 1). We hypothesise that this would be an intrinsic preference for enmeshment which should manifest not only in choices but in how players experience the interaction. To test this, we measure subjective closeness after each round using the inclusion of self in other tool (Baader et al., 2024). It lets players bring two circles to either be distinct or fully overlapped, the degree of overlap represents how close one feels to the other. If collaborating and being independent are truly equivalent for participants, as the theories suggest, reported closeness should not differ between outcomes. However, if people derive intrinsic value from enmeshment, despite identical payoffs, we expect participants to report feeling closer to their partner following mutual cooperation than when remaining independent. This allows us to ask not only whether people choose to enmesh, but whether enmeshment feels meaningfully different from mere coexistence.

Table 1: **Predictions for each theory for each symmetric strategy**. Both player either independent or coordinate, then joint take, stay passive or give. As team reasoning (TR) and virtual bargaining (VB) make equivalent predictions, we show them as single row of the table. The third row shows predictions for our theory of enmeshment. The last row shows dyads' choice frequency in the first block of each game in the first round, then summed over all 8 rounds

	Prisoners' dilemma (PD)							Friends' dilemma (FD)						
Step 1	Independent			Cooperate			Other	Independent			Cooperate			Other
Step 2	T	P	G	T	P	G		T	P	G	T	P	G	
Nash	✓	×	×	×	×	×	×	✓	×	×	×	×	×	×
VB / TB	×	×	×	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	×
Enmesh.	×	×	×	×	✓	×	×	×	✓	×	×	✓	×	×
Dyads (%)	.07	.06	.02	.00	**.26**	.04	.55	.06	**.22**	.03	0.	**.18**	.08	.43

Experiment

Methods

Design The study was a mixed between and within subject design. The between factor was whether dyads started with 8 rounds of Prisoners' dilemma followed by 8 rounds of Friends' dilemma, *PD first*, or 8 rounds of Friends' dilemma followed by 8 rounds of the Prisoners' dilemma, *FD first*. The data and analysis scripts are available at https://github.com/Vbtesh/cogsci2026_lets_be_friends.

Participants We recruited 66 dyads, 132 participants, 66 men and 66 women, on Prolific. The mean age was 43 ± 12. The median time to complete was 26 minutes. Participants were paid \$12 per hours plus an additional \$.02 per bread they managed to bake.

Materials and procedure The experiment was built using the Otree experiment builder (D. L. Chen et al., 2016). After a brief introduction describing the causal rules of baking (Figure 1), participants were assigned a character (Red or Green) and a preferred ingredient (flour or water, counterbalanced across partners). They were told each bread was worth \$0.02 and given tables explaining how bread production depended on ingredient combinations (Figure 3 A and B). Each dyad completed two blocks of 8 rounds, one block of FD and one of PD, with order counterbalanced across dyads. The games differed only in the ingredient sizes available: in FD all ingredients were medium sized; in PD, separating meant bringing one small and one large ingredient (Figure 1 C, D). Strategies were not labelled (e.g., "independent" or "cooperate") to avoid moral bias. Each step was followed by a short animation and a summary of choices and outcomes (Figure 3 A). After each round, participants reported how close they felt to their partner (Baader et al., 2024). At the end of each block, they described their partner in a text box and reported trust on a slider. Before finishing, participants answered three comprehension questions about the causal rules of baking (e.g., “How many breads would you bake with one large water and one small flour?”).

Results

We split our analyses into two parts with distinct goals and ways to explore the data. The first section deals with game theoretic analyses and thus focuses on the difference between PD and FD in the first block. It considers actions across both steps and all trials at once. The second part is interested in dynamics and how different behaviours relate to closeness rating. For simplicity, we only look at actions from the first steps but consider all block and all trials.

Analysis of equilibria

Participants do neither Nash nor virtual bargaining As Table 1 shows, the distributions of participants’ actions shows a systematic pattern of response. In PD, 26% of actions were coordinate passive (*CP*), while in FD, 22% of actions were independent passive (*DP*) and 18% *CP*. Nash clearly fails to explain this pattern as only 7% and 6% of choices in the PD and FD respectively were at the Nash equilibria. Moreover, virtual bargaining and team reasoning are silent about this preference for choosing passive in the second step. By assigning every symmetric strategy equal probability, virtual bargaining and team reasoning can represent choosing to either independent or cooperate in FD but give no insight as to what motivated dyads to take the risk to cooperate. To elucidate the choice to stay passive, we develop a measure for signalling trust through codependency as a way to break ties in virtual bargaining.

Enmeshment as joint dependency in payoffs The empirical equilibria leave a puzzle. Virtual bargaining and team reasoning can account for cooperation in PD, but neither explains why participants systematically choose to remain passive in step two rather than take, even though taking is risk-dominant and increases one’s payoff whenever the other does not also take. Standard accounts of fairness are similarly silent: taking and staying passive can both lead to equal final payoffs depending on the other’s choice. We propose that what participants seek is not merely equal payoffs but *balanced vulnerability*, a state in which both players are meaningfully exposed to each other’s choices. Enmeshment is relational at its core, it does not minimise risk, as Nash predicts, nor maximises joint payoff, as team reasoning suggests, but finds equilibria whose stability stems from the joint knowledge that each player could harm the other but chooses not to. To formalise this, we define a measure of codependency that can serve as a tie-breaker among Pareto-optimal outcomes:

$$\mathrm{Enmesh}(s_i, s_j) = \mathrm{Dep}_i(s_i, s_j) + \mathrm{Dep}_j(s_j, s_i) \tag{1}$$

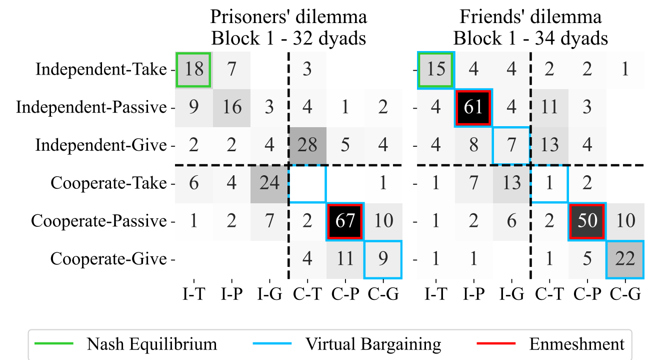

Figure 4: **Frequencies of joint actions across FD and PD.** Number of each joint strategies played by all dyads across all 8 first round of either Prisoners or Friends’ dilemma along with predictions for each competing theory.

where $\mathrm{Dep}_i(s_i, s_j)$ is defined as:

$$\mathrm{Dep}_i(s_i,s_j) = \min\left(\underbrace{u_i(s_i,s_j) - \min_{s'_j}\left(u_i(s_i,s'_j)\right)}_{\text{Vulnerability}}, \underbrace{\max_{s'_j}\left(u_i(s_i,s'_j)\right) - u_i(s_i,s_j)}_{\text{Opportunity}}\right) \tag{2}$$

i.e. the minimum of what player i stands to lose should player j deviate and what they stand to gain under a different choice of j. We take the minimum to reflect our interest in a balanced strategy, neither too vulnerable nor too opportunistic; the sum across players ensures that both must be meaningfully exposed. The crux is that $\mathrm{Enmesh}(s_i, s_j)$ cares about both Dep_i and Dep_j being as high as possible: a state which maximises the vulnerability and opportunity of both players, offering balanced codependency. The selected strategy is thereforeThe selected strategy is therefore $(s_i^*, s_j^*) = \arg\max_{(s_i,s_j)\in P(S)} \mathrm{Enmesh}(s_i, s_j)$, where $P(S)$ denotes the Pareto-optimal outcomes of the strategy set S. The predicted outcome is thus one of mutual engagement: symmetric stakes that signal trust through chosen vulnerability. This account does not, on its own, predict whether participants will cooperate or be independent in step one of FD as those equilibria still score equally on codependency, but it does predict the choice to remain passive in step two, which creates the balanced interdependence that taking or giving would break. We therefore view this formalisation not as a complete theory of enmeshment, but as a first step toward capturing the relational preferences that standard accounts overlook.

Does enmeshment account for participants choices? Our theory of enmeshment predicts three most likely strategies: coordinate-passive (CP) in PD, and both independent-passive (IP) and coordinate-passive in FD. Pooling first-block choices from each condition, Table 1 shows participants’ response distributions peak at these same three points. To formally compare models, we compute the likelihood of participants’

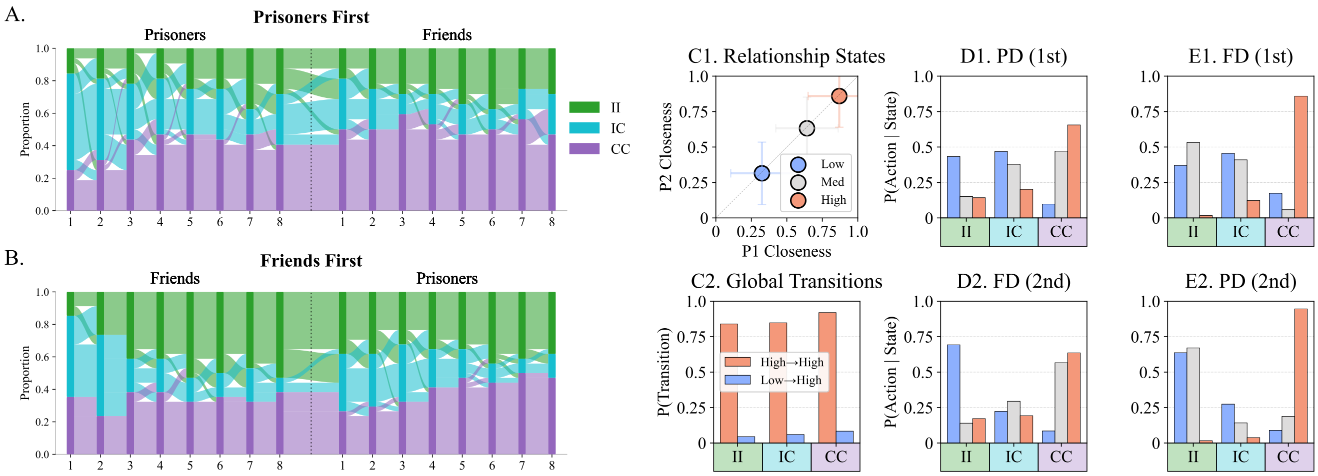

Figure 5: **Trajectories of dyads across both games and conditions and estimated probabilities of each joint action given each latent relationship state**. A. Prisoners' dilemma first; B. Friends' dilemma first. C1. Inferred latent relationship states; C2. Transitions between relationships states; D. Probability of each joint action under each relationship state for *PD first* dyads in 1. PD and 2. FD; E. Probability of each joint action under each relationship state for *FD first* dyads in 1. FD and 2. PD.

first-block data under each theory, assuming dyads noisily drew from categorical distributions parametrised by the predicted equilibria. Each theory is treated as myopic, that is choosing independent or cooperation without considering what could be gained in following the give and take step. We do so by assuming that dyads were noisily drawing from categorical distributions parametrised by the predicted equilibria. We fit an inverse temperature τ to each model. The results in Table 2 confirm that enmeshment best models dyads' choices, with the fit shown in Figure 4. By augmenting virtual bargaining, our measure isolates the contribution of enmeshment in capturing participants' judgments: it predicts the otherwise puzzling choice to stay idle and vulnerable in the second stage. Yet, while enmeshment equally predicts independent and cooperation in FD, which bares out in the data, it remains that collaborating is still more risky and a choice which remains rationally elusive. The match between our model and empirical choices made by dyads can be seen in Figure 4.

Table 2: **Model fits for each theory**

Model $\mathcal{M}$	τ	$-\log p(X\|\mathcal{M})$	Weight
Nash	0.46	2255	0.0
VB / TR	1.70	2098	0.0
Enmeshment	2.46	**2015**	**1.0**

Alternative utility functions A natural objection is that the choice to stay passive might be explained by existing models of social preferences rather than a novel preference for enmeshment. We tested two candidates. First, Fehr-Schmidt inequity aversion, which penalises unequal outcomes asymmetrically depending on who is ahead (Fehr & Schmidt, 1999), $U(R_s, R_o) = R_s + \alpha \max(0, R_s - R_o) + \beta \max(0, R_o - R_s)$, where R_s is the payoff one's gets and R_o is the other's payoff. Second, a model combining adopted utility (valuing the other's payoff) with absolute inequity aversion, $U(R_s, R_o) = R_s + \beta R_o + \delta |R_o - R_s|$. We performed a grid search over plausible parameter ranges to find weights under which cooperate-passive emerges as a Nash equilibrium. For Fehr-Schmidt, no parameterisation produced an equilibrium with any probability mass on cooperate-passive. For the adopted utility model, cooperate-passive appeared only as part of a mixed equilibrium, requiring that players value the other's outcome positively but less than their own, and exhibit strong inequity aversion. This latter finding is intuitive: collaborating requires some concern for the other, but staying passive, rather than giving, requires aversion to unequal outcomes. However, mixed equilibria predict stochastic switching between strategies across trials, whereas our participants showed stable, consistent choices when collaborating. Indeed, the likelihood to keep collaborating after *CP* was .89 in FD and .85 in PD, while the maximum predicted by the inequity aversion utility function is .44 (see Figure 5 A and B where few strategy switches occur between subsequent rounds). The pattern we observe is thus not well captured by these utility functions: they either fail to predict passivity at all, or predict a variability that does not match the data.

Closeness and cooperation across games

Our theory of enmeshment does not provide a clear reason for choosing to cooperate in FD over being independent: both equilibria are equivalent on material and codependency grounds. Yet intuitively, putting oneself at risk by collaborating feels different from acting independently. If cooperation in FD reflects a desire for enmeshment rather than strategic calculation, it should be associated with greater subjective closeness.

We measured closeness after each round using the Inclusion of Other in Self scale (Baader et al., 2024). Within dyads rating correlated as trials progressed with a slight dip at round

8 as some individuals changed behaviour in the last round (first round: $r = -.02, p = .84$; fourth: $r = .39, p < .01$; seventh: $r = .46, p < .001$; eighth $r = .39, p < .01$; sixteenth $r = .38, p < .01$). This suggests that dyad acquired shared representations of the relationship state over both games. A linear mixed effect model predicting closeness from games, (PD, FD), and joint action for step 1: (II, IC, or CC), showed that dyads who cooperated felt closer than those who did not ($\beta = .14 \pm .02, p < .001, z = 8.4$, 95%CI $= [.11, .17]$). There was no difference between PD and FD ($z = .53, p = .6$) and no significant interaction ($z = .75, p = .46$), suggesting that the difference in closeness was due to the *experience* of cooperation rather than the payoff difference resulting from it.

Cooperation signals closeness in FD but not PD. To formally examine how closeness and actions interact differently across games, we fit a hidden Markov model with $K = 3$ latent relationship states (low, moderate, high closeness). States emit both a joint closeness, which are the observed closeness judgment for the dyad after each round (assumed to be Gaussian), and a joint action (categorical and in $\{II, IC, CC\}$), with the emission probabilities for actions varying hierarchically by game and block order. Closeness relationship states transition as a function of the previous state and action taken. We set $K = 3$ latent states with an ordinal constraint such that higher states correspond to greater closeness. Transition probabilities varied by game but not block order. Parameters were estimated using MCMC in NumPyro (Phan et al., 2019). The model converged with maximum $\hat{R} = 1.02$.

The emission probabilities $p(\text{action}|\text{state})$ reveal different patterns across games. In FD, cooperation was almost exclusively associated with the highest closeness state: $p(CC|\text{high}) = .86$, compared to $p(CC|\text{medium}) = .06$ (Figure 5 E1.). In PD, by contrast, cooperation was likely for both medium and high closeness states (.47 vs .66, see Figure 5 D1.). This suggests that dyads perceived cooperation in PD as a noisier signal reflecting either closeness or the selfish incentive to coordinate on the higher payoff. The intermediate state also differed between games. In PD, moderate closeness predicted cooperation (.47) or miscoordination (.38), consistent with attempted cooperation that sometimes fails. In FD, moderate closeness predicted independence (.46) or miscoordination (.41), interpretable as peaceful coexistence: bakers acting independently but not at each other's expense. Low closeness looked similar across both games, with roughly equal probability of choosing independence or miscoordinating.

Across both games, transitions between states were rare (Figure 5 C2.), indicating that relational bonds were stable within the timescale of the experiment. This stability limits our ability to make directional claims about whether actions drive closeness or vice versa. Crucially, the game played first shaped the relationship between actions and closeness for the rest of the experiment (Figure 5 D2. and E2.): dyads who started with PD, where cooperation is strategically ambiguous, maintained a weaker link between cooperation and closeness even when they later played FD, while dyads who started with FD, where cooperation can only signal relational investment, preserved this tighter coupling when they moved to PD.

Discussion and conclusion

We find evidence that people seek enmeshment for its own sake, aligning with theories of joint intentionality (Tomasello, 2008). Participants chose mutual vulnerability over safety, risking payoffs to signal trust (Costa-Gomes et al., 2014). Neither Nash, virtual bargaining, nor team reasoning predicted this pattern. Explaining it required augmenting virtual bargaining with an intrinsic drive for balanced codependency: a preference for states where both players are meaningfully exposed to each other's choices, communicating a desire for equality over and beyond equal payoffs (A. M. Chen & Saxe, 2024).

Yet our formalisation remains incomplete. It correctly predicts the choice to stay passive rather than take, but treats cooperation and independence in FD as equivalent, though participants do not. Cooperation in FD was almost exclusively associated with high closeness, while independence was compatible with moderate closeness; dyads distinguished between these equilibria even though material payoffs and codependency scores did not. Something beyond balanced vulnerability governs the felt difference between doing things together and merely coexisting. The link between cooperation and closeness was also stronger in FD than in PD: where cooperation is strategically rational, it provides a noisier signal of relational investment, but where it offers no material advantage, it can only mean one thing. Perhaps most notably, the game played first established a meaning for cooperation that persisted into the second, shaping how actions were interpreted as relational signals.

The central challenge is now to explain why cooperation feels different from independence even when payoffs are identical. We believe Theory of Mind will be paramount (Ho et al., 2022): models of mental state inference allow agents to represent not only what others will do, but what they believe, want, and expect, and to define goals over those (Btesh et al., 2025; Ho et al., 2021; Tejwani et al., 2022). This machinery could capture what our formalisation misses: that cooperating reflects aligned intentions, that choosing interdependence signals a willingness to be known, and that relationships are built through mutual recognition of shared interests and goals (Brockbank et al., 2025; Mahaphanit et al., 2025). The question is no longer whether people prefer enmeshment, but how minds represent the relational meaning of joint action and map it to different kinds of relationships (Davis et al., 2026).

Acknowledgements

VB was supported by a Bogue Fellowship and a UCL Experimental Psychology Demonstratorship.

References

Baader, M., Starmer, C., Tufano, F., & Gächter, S. (2024). Introducing IOS11 as an extended interactive version of the 'Inclusion of Other in the Self' scale to estimate relationship closeness. *Scientific Reports*, *14*(1).

Becker, G. S. (1976). *The Economic Approach to Human Behavior*. University of Chicago Press.

Brockbank, E., Gerstenberg, T., Fan, J. E., & Hawkins, R. (2025). How do we get to know someone? Diagnostic questions for inferring personal traits. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Btesh, V., Lagnado, D., & Gerstenberg, T. (2025). Taking others for granted: Balancing personal and presentational goals in action selection. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Chater, N., Zeitoun, H., & Melkonyan, T. (2022). The paradox of social interaction: Shared intentionality, we-reasoning, and virtual bargaining. *Psychological Review*, *129*(3), 415–437.

Chen, A. M., & Saxe, R. (2024). How taking turns communicates desired equality in social relationships. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*(100), 88–97.

Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences*, *26*(2), 139–153.

Colman, A. M., & Gold, N. (2018). Team reasoning: Solving the puzzle of coordination. *Psychonomic Bulletin & Review*, *25*(5), 1770–1783.

Colman, A. M., Pulford, B. D., & Rose, J. (2008). Team reasoning and collective rationality: Piercing the veil of obviousness. *Acta Psychologica*, *128*(2), 409–412.

Costa-Gomes, M. A., Huck, S., & Weizsäcker, G. (2014). Beliefs and actions in the trust game: Creating instrumental variables to estimate the causal effect. *Games and Economic Behavior*, *88*, 298–309.

Davis, I., Jara-Ettinger, J., & Dunham, Y. (2026). Inferring the internal structure of groups through the integration of statistical learning and causal reasoning. *Nature Communications*, *17*(1).

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

Harsanyi, J., & Selten, R. (1988). *A General Theory of Equilibrium in Games*. The MIT Press.

Hilbe, C., Schmid, L., Tkadlec, J., Chatterjee, K., & Nowak, M. A. (2018). Indirect reciprocity with private, noisy, and incomplete information. *Proceedings of the National Academy of Sciences*, *115*(48), 12241–12246.

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2021). Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General*, *150*(11), 2246–2272.

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

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).

Le Pargneux, A., Chater, N., & Zeitoun, H. (2024). Contractualist tendencies and reasoning in moral judgment and decision making. *Cognition*, *249*.

Mahaphanit, W., Keller, B., Hawkins, R., Phillips, J., & Chang, L. (2025). Learning to communicate a shared wavelength. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291–1298.

Phan, D., Pradhan, N., & Jankowiak, M. (2019). Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. *arXiv preprint arXiv:1912.11554*.

Rapoport, A., & Chammah, A. M. (1965). *Prisoner's Dilemma: A Study in Conflict and Cooperation*. University of Michigan Press.

Santos, F. P., Pacheco, J. M., & Santos, F. C. (2021). The complexity of human cooperation under indirect reciprocity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1838).

Schnell, E., & Muthukrishna, M. (2024). Indirect reciprocity undermines indirect reciprocity destabilizing large-scale cooperation. *Proceedings of the National Academy of Sciences*, *121*(19).

Skyrms, B. (2003). *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press.

Tejwani, R., Kuo, Y.-L., Shu, T., Katz, B., & Barbu, A. (2022, November). Social Interactions as Recursive MDPs. In A. Faust, D. Hsu, & G. Neumann (Eds.), *Proceedings of the 5th Conference on Robot Learning* (pp. 949–958, Vol. 164).

Tomasello, M. (2008). *Origins of human communication*. MIT press.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior* [Pages: xviii, 625]. Princeton University Press.

Warneken, F., Chen, F., & Tomasello, M. (2006). Cooperative Activities in Young Children and Chimpanzees. *Child Development*, *77*(3), 640–663.

Warneken, F., Gräfenhain, M., & Tomasello, M. (2012). Collaborative partner or social tool? New evidence for young children's understanding of joint intentions in collaborative activities. *Developmental Science*, *15*(1), 54–61.

Warneken, F., & Tomasello, M. (2006). Altruistic Helping in Human Infants and Young Chimpanzees. *Science*, *311*(5765), 1301–1303.