Inferring Joint Commitment from Observed Coordination

Jinyi Kuang[1] **(jkuang@sas.upenn.edu), Erik Brockbank**[2]**, Cristina Bicchieri**[1]**, Robert D. Hawkins**[2]
[1]University of Pennsylvania [2]Stanford University

Abstract

Successful social life requires agents to accurately perceive social bonds and infer whether interacting partners are committed to a shared goal. Yet observable behavior is often ambiguous regarding whether a joint commitment is present. How do observers infer joint commitment from observed behavior alone? We propose a computational model of commitment inference based on Bayesian Theory of Mind. The model predicts that observers update beliefs about latent commitments based on two cues: whether coordination requires mutual participation to avoid a suboptimal outcome (interdependence) and the history of successful coordination (repetition). We tested these predictions in two experiments ($N = 785$) in which participants observed two farmers harvesting berries in a grid world farm and evaluated the level of commitment between them. Results show that both interdependence and repetition strengthen inferences of commitment. These findings demonstrate how observers track the emergence of collective agency from observed coordination.

Keywords: joint action; joint commitment; Bayesian theory of mind; coordination

Introduction

Successful social life depends on the ability to coordinate with others. From carrying furniture to playing music together, people routinely align their behavior to pursue shared goals. In many such cases, coordination relies on joint commitment: agents acting together as a unit toward a shared goal. Recognizing joint commitment matters not only for the interacting agents, but also for observers. When observers perceive a joint commitment, they can better predict future coordination, explain why agents remain engaged in a shared task (Michael et al., 2016b), and enforce social norms by holding agents accountable for deviation (Gilbert, 1999). Yet observable behavior is often ambiguous with respect to shared goals. Two individuals may walk side by side because they are going somewhere together, or simply because they happen to be heading in the same direction. Because joint commitment is not always made explicit in everyday interactions, either because doing so is unnecessary or because coordination can proceed without negotiation, particularly in familiar or low-stakes contexts (Clark, 2020), observers cannot rely on explicit signals alone. How, then, do they infer the presence of joint commitment from observed actions?

Existing psychological research suggests that observers rely on a range of cues and heuristics to resolve this ambiguity, including prior interaction history (Michael et al., 2016a), the effortfulness of the activity (Bonalumi et al., 2019; Michael & Pacherie, 2015), minimal communicative signals such as eye contact or brief gestures (Bangerter et al., 2022; Siposova et al., 2018), and the degree of social dependency between agents (Bonalumi et al., 2022; Michael et al., 2016b). However, these accounts remain largely descriptive: they identify which cues influence judgments of commitment, but do not specify the underlying inference process that generates those judgments. In particular, they do not formalize what it means for an observer to treat commitment as a hidden property of an interaction, nor how such a property could be inferred from noisy, temporally extended behavior.

In the present work, we address this gap by modeling joint commitment as a latent state that observers infer from observing repeated social interactions. We build on Bayesian Theory of Mind (BToM) approaches, which model how agents infer others' hidden mental states from observed behavior (Baker et al., 2017) while solving coordination problems (Kleiman-Weiner et al., 2025; Meulemans et al., 2025). Extending these frameworks, we propose that observers infer joint commitment by tracking the extent to which agents' actions reflect a shared decision process, appearing to act as a "we" agent, rather than independent goal pursuit. This notion draws on philosophical and computational accounts of collective intentionality and team reasoning, which characterize joint action in terms of agents adopting group-directed goals or optimizing joint outcomes (Bacharach, 2006; Gilbert, 1999; Gold et al., 2007; Searle, 1990; Stacy et al., 2024; Tang et al., 2025; Tuomela & Miller, 1988). Importantly, we use "acting as a we" as an observer-level characterization of coordination under joint commitment, without committing to a particular account of collective intentionality.

Our model captures how evidence accumulates over time from patterns of interdependence and spatiotemporal coordination, which allow observers to infer graded levels of commitment. To evaluate this account, we introduce a grid world environment in which two agents engage in joint action under controlled variations of repetition, co-location, and payoff interdependence. Across two experiments ($N = 785$), participants observed agents harvesting berries and judged their level of commitment. Results show that observers integrate both social-structural cues and visible coordination patterns, supporting a computational account of how joint commitment is inferred from action.

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

Computational Model

Inference from Risky Coordination

We begin by specifying how commitment shapes action. Agents can act under two perspectives. Under an *individual* ("I") perspective, agents evaluate outcomes based on their own payoffs and face uncertainty about their partner's behavior. Under a *collective* ("We") perspective, agents behave as if actions are generated by a shared decision process which eliminates uncertainty about coordination. Let p denote the agent's belief that their partner will coordinate, and let payoffs be R_H (successful coordination), R_M (safe option), and R_L (failed coordination). Table 1 shows the expected utilities under each perspective:

Perspective	$U(\text{coord})$	$U(\text{defect})$
"I"	$p \cdot R_H + (1-p) \cdot R_L$	R_M
"We"	R_H	R_M

Table 1: Individual and collective expected utilities.

We define a continuous commitment parameter $\theta \in [0,1]$, which captures the extent to which observed behavior is explained by the collective perspective rather than the individual perspective. Expected utility is modeled as a weighted combination of these two perspectives:

$$\Delta U(\theta) = (1-\theta) \cdot \Delta U_I + \theta \cdot \Delta U_{We} \tag{1}$$

Agents select actions using a softmax over expected utility:

$$P(\text{coord} \mid \theta) = \sigma\big(\beta \cdot \Delta U(\theta)\big) \tag{2}$$

with a logistic function $\sigma(\cdot)$. β reflects choice stochasticity. Observers do not directly observe θ, but infer it from agents' behavior over repeated interactions T using Bayes' rule:

$$P(\theta \mid T) \propto P(\text{coord} \mid \theta)^T \cdot P(\theta) \tag{3}$$

This model assumes that, when payoffs are interdependent, agents acting from an individual perspective ($\theta = 0$) face uncertainty about their partner (e.g., an uninformative belief such as $p \approx 0.5$), which makes coordination risky because it yields R_H only if both coordinate and R_L otherwise. In this case, agents with low θ are less likely to coordinate, but become increasingly likely to coordinate as θ approaches 1, which reflects a shift toward a joint perspective that eliminates partner uncertainty. When payoffs are independent, such that coordination yields R_H regardless of the partner's action, $P(\text{coord} \mid \theta)$ remains high across values of θ. Consequently, when coordination is risky, observing agents repeatedly coordinate provides increasingly strong evidence for joint commitment (higher θ) from an observer's perspective.

Inference from Spatiotemporal Co-location

Commitment can also be inferred from sustained engagement in synchronized activity. Prior work suggests that agents who move together or remain aligned are perceived as a coherent unit (i.e., "entitativity") (Campbell, 1958; Waytz & Young, 2012). We model this as a heuristic in which observers treat consistent spatial alignment as probabilistic evidence about the same latent commitment parameter θ, regardless of the underlying payoff structure. Co-location is assumed to arise from a mixture of two processes: (i) a commitment-based process, in which proximity reflects an underlying shared structure, and (ii) an uninformative process, in which proximity is incidental. A reliability parameter $\lambda \in [0,1]$ determines the weight assigned to co-location as an informative signal.

$$P(\text{together} \mid \theta) = \lambda\big[\theta + (1-\theta)\varepsilon\big] + (1-\lambda) \cdot \tfrac{1}{2}, \tag{4}$$

where $\varepsilon \ll 1$ captures accidental co-location under low commitment, and $\frac{1}{2}$ is an uninformative baseline. When $\lambda = 1$, co-location is fully trusted as evidence for commitment; when $\lambda = 0$, it is ignored. This formulation treats spatial alignment as a probabilistic cue whose influence depends on its perceived reliability, consistent with Bayesian accounts of cue integration (Ernst & Banks, 2002).

Taken together, observers infer the latent commitment parameter θ by integrating two distinct sources of evidence over T rounds: *social reasoning* about coordination under risk (Equation 2) and the *heuristic* of visual co-location (Equation 4). The posterior distribution is:

$$P(\theta \mid T) \propto \underbrace{P(\text{coord} \mid \theta)^T}_{\text{social reasoning}} \cdot \underbrace{P(\text{together} \mid \theta)^T}_{\text{heuristic}} \cdot P(\theta) \tag{5}$$

We test this model in two experiments that systematically manipulate the observable evidence for joint commitment from social reasoning and spatiotemporal co-location.

Experiments

All Experiments from the current study were preregistered on AsPredicted. Experiment code and analyses are available on GitHub at https://tinyurl.com/gridfarm.

Participants We recruited 800 participants from Prolific across Experiments 1 and 2. All participants used English as their primary language, resided in the U.S., and had a completion rate of at least 95%. Out of 400 participants in Experiment 1, one was excluded due to technical error, resulting in a final sample of $N = 399$ (Female: $N = 199$, 49.9%, age: $M(SD) = 42.5(12.7)$). Participants were paid \$1.25 for completion (estimated duration: 5 minutes; median completion time: 4.43 minutes). Out of 400 participants recruited for Experiment 2, two were excluded due to technical error and 12 were excluded for failing an attention check, resulting in a final sample of $N = 386$ participants (Female: $N = 196$, 50.8%, age: $M(SD) = 41.8(12.4)$). Participants were paid \$1.25 for completion (estimated duration: 5 minutes; median completion time: 5.32 minutes).

Figure 1: Task environments in Experiments 1 and 2. Participants observed repeated coordination (2 vs. 6 rounds) between two agents in a 10 x 10 grid world, followed by a defection trial. Blue text indicates payoffs in the *interdependent* conditions, and red text indicates payoffs in the *independent* conditions.

Design Both experiments used a 2 × 2 between-subjects factorial design. The independent variables were repetition of observed successful coordination (*low repetition*: 2 rounds vs. *high repetition*: 6 rounds) and payoff interdependence (*interdependent* vs. *independent*). Participants were randomly assigned to one of the four conditions.

Materials The task environment consisted of a 10×10 grid containing two agents, Yellow Farmer and Purple Farmer, that began each trial at diagonal corners of the grid (Figure 1). In Experiment 1, the grid contained one center tree and two corner trees offset from the farmers' respective corners. In the *interdependent* condition, the payoff structure instantiated a Stag Hunt game: the tree in the center yielded 8 berries per farmer, but only if both farmers harvested it together. If one farmer harvested the center tree, they received just 1 berry because the branches were too high for one farmer to reach alone. Smaller trees in the corners yielded 5 berries and could be harvested by either farmer on their own. In the *independent* condition, a farmer who harvested the center tree would receive 8 berries regardless of the other farmer's action.

In Experiment 2, the grid contained two berry trees, one closer to each farmer's start location. The tree closer to Yellow yielded 5 yellow berries and 8 purple berries, whereas the tree closer to Purple yielded 8 yellow berries and 5 purple berries. If both farmers harvested the tree farther from their respective start locations, each received 8 berries; if both harvested the closer tree, each received 5 berries. In the *interdependent* condition, if one farmer harvested the tree closer to their start location and the other traveled to the farther tree, the one that arrived earlier would receive 5 berries, but this action interfered with the farmer who traveled farther and arrived later, reducing the second farmer's payoff to 1 berry. In the *independent* condition, one farmer's harvesting decision did not affect the other's outcome; the one who traveled farther continued to receive 8 berries regardless of the other farmer's choice. As in Experiment 1, the *interdependent* condition instantiated a Stag Hunt payoff structure, but the payoff-dominant coordination solution required farmers to go their separate ways rather than meeting together. This addresses a potential alternative interpretation of the coordination behavior in Experiment 1. There, when agents converge on the same location, observers may assume explicit communication. In Experiment 2, by having agents harvest different locations as a form of cooperation, the possibility of explicit communication between them is less straightforward.

Procedure The procedure was identical across experiments. Participants were instructed that they would observe several harvests in which two farmers harvested berries from a shared farm. Participants learned all possible actions each farmer could take and the corresponding outcomes. A comprehension check following the instructions asked participants several multiple-choice questions about the payoff structure. Participants could not continue until they had answered all questions correctly.

During the task, participants first completed an observation phase, during which they observed either 2 rounds (*low repetition*) or 6 rounds (*high repetition*) in which both farmers successfully coordinated and received the payoff-maximizing outcome (8 berries each). Participants observed the farmers' harvest paths and resulting payoffs in each round. Following the observation phase, participants observed a critical trial. In this trial, Purple Farmer deviated by harvesting a lower-payoff option (receiving 5 berries), while Yellow Farmer pursued the higher-payoff option, expecting Purple's cooperation. As a result, Yellow received either 1 berry in the *interdependent* condition or 8 berries in the *independent* condition.

Measures After the critical trial, participants answered four sliding scale questions ranging from 0 to 100, with 0 representing *not at all* and 100 representing *extremely*: 1) *Agreement*: "To what extent do you think Yellow and Purple had an unspoken agreement to [harvest the center tree together][harvest the farther tree separately]?"; 2) *Commitment*: "To what extent do you think Yellow and Purple are committed to [harvest the center tree together][harvest the farther tree separately]?"; 3) *Anger*: "How angry would Yellow feel that Purple went to [the corner tree][the tree near them]?"; 4) *Guilt*: "How guilty would Purple feel about going to [the corner tree][the tree near them]?" The agreement and commitment measures assess perceived joint commitment directly. The anger and guilt measures capture normative reactions to commitment deviations (Castro & Pacherie, 2022). We conceptualize normativity in our task as social rather than moral, as deviations are more likely experienced as violations of interpersonal expectation than breaches of broader moral principles (see Turiel, 1983). Framing such deviations as "moral" may impose an overly strong normative lens. We collected these measures only once, at the end of the sequence, rather than trial-by-trial because repeated probing during the interaction might encourage explicit monitoring of the measures.

Analysis For each dependent variable (Agreement, Commitment, Anger, and Guilt), we conducted a two-way between-subjects analysis of variance (ANOVA) with repetition condition (*low* vs *high*) and interdependence (*independent* vs *interdependent*) as fixed factors. For each ANOVA, we first tested the interaction between repetition and interdependence. When the interaction was not significant, the interaction term was not interpreted further, and only the main effects of repetition and interdependence are reported. We conducted similar ANOVA analyses with pooled data from each experiment, including the experiment as a predictor.

Results

Behavioral results

Repetition and Payoff Interdependence Predict Commitment Across both experiments, participants inferred significantly higher agreement and joint commitment in the *high repetition* condition than the *low repetition* condition (Figure 2). While interdependence increased both Agreement and Commitment inferences in Experiment 1, this effect was only significant for Agreement inferences in Experiment 2. We found no significant interactions between repetition and interdependence for any measures (all $ps > .16$). Anger and Guilt attributions were significantly higher in the *interdependent* condition (see Table 2 for full statistics).

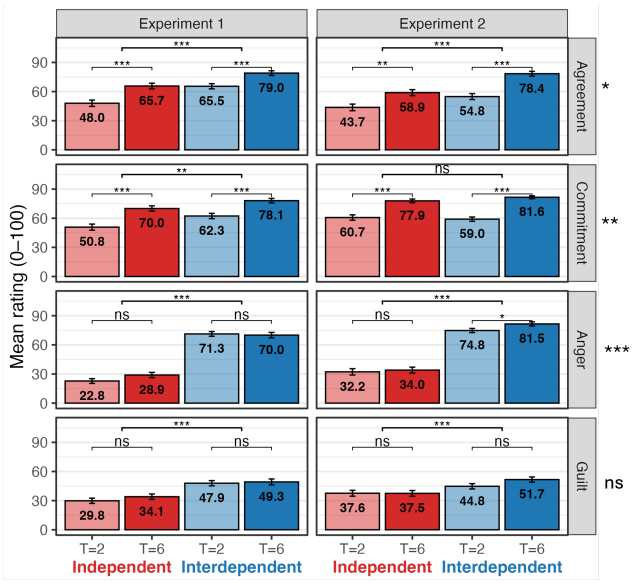

Figure 2: Results from Experiments 1 and 2. The bar charts show the mean ratings (on a scale of 0–100) for Agreement, Commitment, Anger, and Guilt. Error bars indicate the standard error of the mean. Symbols outside the panels indicate statistical comparisons between experiments. Statistical significance between groups is marked by brackets and asterisks ($*p < .05, **p < .01, ***p < .001$, 'ns' not significant).

Inferred commitment predicts normative reactions to defection To examine whether inferred joint commitment predicts normative reactions to defection, we conducted exploratory linear regression analyses predicting Anger and Guilt from observers' judgments of Agreement and Commitment. All models include payoff interdependence, repetition condition, and visual co-location (i.e., experiment) as covariates. We found that observers' judgments of unspoken agreement and joint commitment predicted both Anger and Guilt, with higher perceived agreement associated with greater anger toward the defector ($\beta = 0.30$, 95% CI $[0.24, 0.36]$, $p < .001$) and greater guilt attributed to the defector ($\beta = 0.30$, 95% CI $[0.24, 0.37]$, $p < .001$). Similarly, inferred joint commitment predicted both Anger ($\beta = 0.16$, 95% CI $[0.09, 0.24]$, $p < .001$) and Guilt ($\beta = 0.25$, 95% CI $[0.17, 0.33]$, $p < .001$).

Spatiotemporal Co-location Enhances Commitment Inference To test the hypothesis that visual co-location serves

as a heuristic for inferred commitment, we compared results across experiments. Exploratory analyses found that ratings for Agreement ($F(1,777) = 6.41,\ p = .012,\ \eta_p^2 = .01$) and Commitment ($F(1,777) = 7.39,\ p = .007,\ \eta_p^2 = .01$) were significantly higher in Experiment 1 (co-located) than in Experiment 2 (separate). Anger was also significantly higher in Experiment 1 ($F(1,777) = 19.50$, $p < .001$, $\eta_p^2 = .02$); Guilt did not differ significantly across experiments ($p = .131$).

Model Prediction

Model fitting We fit the computational model jointly to data from Experiments 1 and 2, with separate fits for Agreement and Commitment. The model predicted a latent commitment estimate θ, mapped onto the 0–100 response scale as $100 \times E[\theta]$. The full model included both the payoff-based and spatiotemporal components, while reduced models removed one or both. Parameters were estimated by maximizing likelihood under a Gaussian observation model. Free parameters included the inverse temperature β, visual reliability λ, and observation noise σ, with constraints $\beta > 0$, $\lambda \in [0,1]$, and $\sigma > 0$. Model predictions were compared to mean participant ratings (Figure 3). The model captured Agreement well ($r = .90$, RMSE = 5.25), but fit for Commitment was weaker ($r = .51$, RMSE = 11.21) which suggests additional factors may influence commitment judgments.

Model comparison We compared the full model to reduced alternatives (social reasoning-only, heuristic-only, null) using 10-fold cross-validation. Predictive performance was assessed via expected log pointwise predictive density (ELPD; see Table 3). For Agreement, the full model provided the best fit (ΔELPD = 0). For Commitment, the null model slightly outperformed the full model, though differences were modest. Cross-validated accuracy showed a similar pattern: the full model performed best for Agreement ($r = .34$, RMSE = 29.49), whereas differences were smaller for Commitment (null RMSE = 26.52 vs. 27.02).

Table 2: Summary of ANOVA results (Main Effects) for Experiments 1 and 2.

	Experiment 1			Experiment 2		
Measure / Effect	$F(1,395)$	p	η_p^2	$F(1,382)$	p	η_p^2
Agreement						
Repetition	26.70	$< .001$	.06	40.50	$< .001$	.10
Interdependence	28.68	$< .001$	.07	26.12	$< .001$	.06
Commitment						
Repetition	37.36	$< .001$	.09	86.96	$< .001$	.19
Interdependence	12.22	.001	.03	0.26	.612	.00
Anger						
Repetition	0.01	.916	.00	2.20	.139	.01
Interdependence	287.77	$< .001$	.42	276.06	$< .001$	.42
Guilt						
Repetition	0.39	.535	.00	1.33	.249	.00
Interdependence	34.46	$< .001$	.08	14.31	$< .001$	.04

Note. No Rep. $\times$ Interdep. interactions were significant.

Table 3: Model comparison based on 10-fold cross-validated ELPD. Standard errors (SE) are computed following Vehtari et al. (2017).

	Agreement		Commitment	
Model	ΔELPD	SE	ΔELPD	SE
Full Model	0.00	—	14.59	—
Social Only	13.92	5.36	40.55	7.60
Heuristic Only	49.83	10.39	79.09	10.80
Null Model	47.64	9.51	0.00	7.52

Parameter recovery We conducted a parameter recovery analysis using simulated datasets matched to the experimental design. Recovery accuracy was assessed by correlating true and recovered parameter values across simulations. The heuristic parameter λ ($r = .945$, RMSE = 0.048) and observation noise σ ($r = .973$, RMSE = 3.48) were well recovered, whereas the inverse temperature parameter β ($r = -.554$, RMSE = 2.39) was less reliably recovered, indicating that the present design provides limited information for distinguishing differences in decision noise.

Residual analysis We examined whether the model systematically over- or under-predicts judgments, as well as overall prediction error as a function of repetition, interdependence, and measure (Agreement vs. Commitment). With increasing repetition, the model tended to underestimate participants' ratings ($\beta = 8.01,\ p < .001$), even though overall prediction error decreased ($\beta = -3.34,\ p = .003$). Prediction error was also lower in interdependent conditions ($\beta = -8.45$, $p < .001$), suggesting improved fit when coordination involved risk. Commitment judgments were less well predicted in the *independent* condition and better predicted in the *interdependent* condition ($\beta = -9.70,\ p < .001$), whereas Agreement predictions were comparable across conditions. Residual variance was unrelated to participant effort ($p = .997$) but decreased with judged difficulty ($\beta = -1.74,\ p < .001$).

General Discussion

The present work investigated how observers infer joint commitment from observed coordination in the absence of explicit communication. Across two experiments, we found that observers relied on both the structural properties of the interaction and its temporal history to infer whether agents were acting as a "we." Repetition of successful coordination increased perceived agreement and commitment, while payoff interdependence strengthened commitment inferences. These inferences carried normative consequences: higher inferred commitment predicted stronger expectations of anger and guilt following defection. This finding is consistent with philosophical and psychological accounts that treat joint commitment as creating a normative bond between agents (Gilbert, 1999). Together, these findings suggest that observers treat coordinated action not merely as behavioral alignment, but as evidence of an implicit joint commitment.

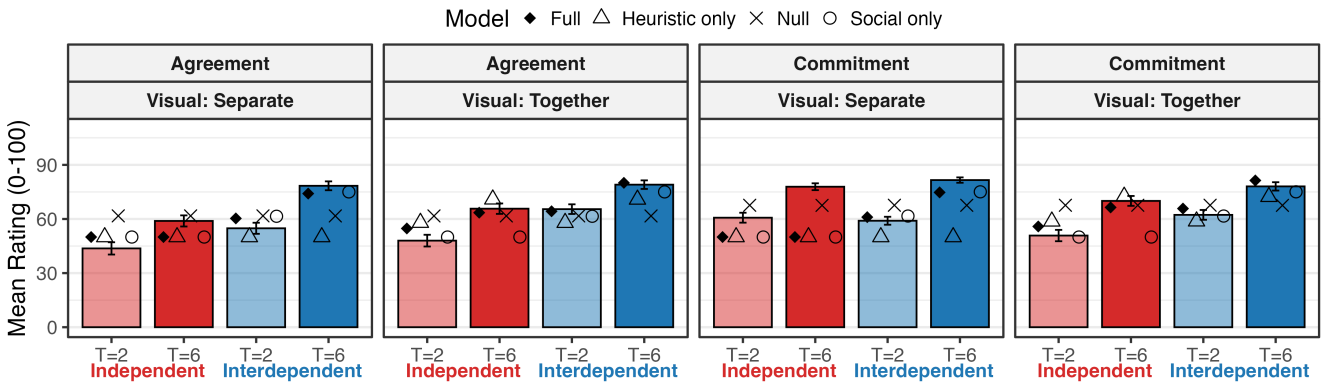

Figure 3: Model predictions against human judgments of Agreement and Commitment. Bars represent mean participant ratings (0–100) of Agreement and Joint Commitment across Experiment 1 (Visual: Together) and Experiment 2 (Visual: Separate). Conditions are stratified by payoff structure (Red = Independent, Blue = Interdependent) and repetition history (T=2 vs. T=6). Error bars indicate standard error of the mean (±SEM). Symbols represent predictions from the full and ablated models.

Our findings align with the predictions of a computational model that combines multiple diagnostic cues to infer joint commitment. When coordination required mutual participation to avoid the risk of a suboptimal outcome, observers treated successful coordination as diagnostic of a shared commitment, consistent with Bayesian Theory of Mind accounts of social inference (Baker et al., 2017). At the same time, repetition influenced agreement and commitment inference even in independent contexts in which coordination was not risky. This pattern is partially explained by a spatiotemporal heuristic in which sustained alignment of behavior can be interpreted as evidence of collective agency (Campbell, 1958) independent of payoff structure. Together, results suggest that observers integrate both social reasoning about payoff structure and heuristic cues about spatiotemporal alignment when inferring joint commitment. More broadly, the present findings speak to how observers interpret coordinated behavior, rather than to the metaphysics of joint action. Philosophical accounts differ in whether joint commitment involves a distinctive normative bond (Gilbert, 1999) or can be explained in terms of interlocking individual intentions and plans (Bratman, 1993). Our results suggest that observers may attribute commitment based on cues such as interdependence and repeated coordination, even when these cues are compatible with weaker forms of shared activity. Thus, the present model should be understood as a theory of commitment inference rather than a theory of joint commitment itself.

We also found that the model predicted agreement judgments better than commitment judgments. One possible explanation is that the model conceptualizes joint commitment primarily as a descriptive property of behavior (i.e., the extent to which agents appear to adopt a shared, interdependent perspective). While agreement judgments may reflect this structure more directly, commitment judgments may reflect additional normative considerations such as obligation and accountability (Gilbert, 1999). Consistent with this interpretation, we found commitment inference was associated with normative reactions to defection, including anger and guilt. Extending the model to incorporate normative aspects may improve its ability to capture commitment judgments.

The present work has several limitations. First, the grid world paradigm captures only a simplified form of social interaction. Although it extends standard matrix games to a spatial setting, it omits many features of coordination in naturalistic environments. Future work could extend this approach to more dynamic settings, such as immersive video games (e.g., Wu et al., 2025), to better capture how agents adapt to one another over time. Second, we did not provide participants with a formal definition of commitment or agreement, as our goal was to capture lay intuitions rather than impose a particular theoretical account. However, this leaves open the possibility that participants interpreted "being committed" in different ways, including both a social or normative sense (e.g., obligation to another agent) and a psychological sense (e.g., persistence or intention toward a course of action). As a result, responses may reflect a mixture of these interpretations. Future work could better characterize lay concepts of commitment by disentangling social obligations from individual intentions. Finally, the present study focuses on spatial coordination as a cue to joint action; however, temporal coordination (e.g., synchrony or rhythm) likely provides an additional and important signal (see Hove and Risen, 2009; Keller, 2008; see also Shamay-Tsoory et al., 2024). Future work could examine how observers integrate spatial and temporal cues when inferring agreement and commitment.

References

Bacharach, M. (2006). *Beyond individual choice: Teams and frames in game theory*. Princeton University Press.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Bangerter, A., Genty, E., Heesen, R., Rossano, F., & Zuberbühler, K. (2022). Every product needs a process: Unpacking joint commitment as a process across species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1859), 20210095.

Bonalumi, F., Isella, M., & Michael, J. (2019). Cueing implicit commitment. *Review of Philosophy and Psychology*, *10*(4), 669–688.

Bonalumi, F., Michael, J., & Heintz, C. (2022). Perceiving commitments: When we both know that you are counting on me. *Mind & Language*, *37*(4), 502–524.

Bratman, M. E. (1993). Shared intention. *Ethics*, *104*(1), 97–113.

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science*, *3*(1), 14–25.

Castro, V. F., & Pacherie, E. (2022). Commitments and the sense of joint agency. *Mind and Language*, *38*(3), 889–906.

Clark, H. H. (2020, August 21). Social actions, social commitments. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of human sociality* (1st ed., pp. 126–150). Routledge.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Gilbert, M. (1999). Obligation and joint commitment. *Utilitas*, *11*(2), 143–163.

Gold, N., Sugden, R., & Journal of Philosophy, Inc. (2007). Collective intentions and team agency: *Journal of Philosophy*, *104*(3), 109–137.

Hove, M. J., & Risen, J. L. (2009). It's all in the timing: Interpersonal synchrony increases affiliation. *Social Cognition*, *27*(6), 949–960.

Keller, P. E. (2008). Joint action in music performance. In F. Morganti, A. Carassa, F. Morganti, & G. Riva (Eds.), *Enacting intersubjectivity. a cognitive and social perspective on the study of interactions* (pp. 205–221).

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.

Meulemans, A., Nasser, R., Wołczyk, M., Weis, M. A., Kobayashi, S., Richards, B., Lajoie, G., Steger, A., Hutter, M., Manyika, J., Saurous, R. A., Sacramento, J., & Arcas, B. A. y. (2025). Embedded universal predictive intelligence: A coherent framework for multi-agent learning.

Michael, J., & Pacherie, E. (2015). On commitments and other uncertainty reduction tools in joint action. *Journal of Social Ontology*, *1*(1), 89–120.

Michael, J., Sebanz, N., & Knoblich, G. (2016a). The sense of commitment: A minimal approach. *Frontiers in Psychology*, *6*.

Michael, J., Sebanz, N., & Knoblich, G. (2016b). Observing joint action: Coordination creates commitment. *Cognition*, *157*, 106–113.

Searle, J. (1990). Collective intentions and actions. In P. R. C. J. Morgan & M. Pollack (Eds.), *Intentions in communication* (pp. 401–415). MIT Press.

Shamay-Tsoory, S. G., Marton-Alper, I. Z., & Markus, A. (2024). Post-interaction neuroplasticity of inter-brain networks underlies the development of social relationship. *iScience*, *27*(2), 108796.

Siposova, B., Tomasello, M., & Carpenter, M. (2018). Communicative eye contact signals a commitment to cooperate for young children. *Cognition*, *179*, 192–201.

Stacy, S., Gong, S., Parab, A., Zhao, M., Jiang, K., & Gao, T. (2024). A bayesian theory of mind approach to modeling cooperation and communication. *WIREs Computational Statistics*, *16*(1), e1631.

Tang, N., Gong, S., Zhao, M., Zhou, J., Shen, M., & Gao, T. (2025). Joint commitment in human cooperative hunting through an 'imagined we'. *Proceedings of the Royal Society B: Biological Sciences*, *292*(2055), 20243070.

Tuomela, R., & Miller, K. (1988). We-intentions. *Philosophical Studies*, *53*(3), 367–389.

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge University Press.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432.

Waytz, A., & Young, L. (2012). The group-member mind trade-off: Attributing mind to groups versus group members. *Psychological Science*, *23*(1), 77–85.

Wu, C. M., Deffner, D., Kahl, B., Meder, B., Ho, M. K., & Kurvers, R. H. J. M. (2025). Adaptive mechanisms of social and asocial learning in immersive collective foraging. *Nature Communications*, *16*(1), 3539.