The Agency Gap: Testing Human–Human and Human–AI Collaboration Using a Minimal Coordination Game

Shaozhe Cheng (shaozhe.cheng@duke.edu)[1,2], Tao Gao[2], Michael Tomasello[1] & Tamar Kushnir[1]
[1]Department of Psychology & Neuroscience, Duke University
[2]Department of Communication, UCLA

Abstract

Humans coordinate by forming and committing to shared intentions—a form of shared agency that is largely absent in standard reinforcement learning agents and remains unclear in large language models, which function more like tools than autonomous agents. We compared human–human and human–AI collaborations using a non-linguistic sequential coordination game designed to disentangle successful collaboration outcomes from the underlying coordination mechanisms, such as commitment to an established joint goal despite the emergence of conflicting alternatives. Although adults and 7-year-old children were often able to adapt to different AI partners and achieve successful outcomes, human–human collaboration was marked by uniquely high levels of commitment and coordination efficiency, both of which were reduced in human–AI interactions. Five-year-old children showed greater difficulty collaborating with AI partners that functioned merely as asocial tools. Our findings suggest that human-compatible AI require mechanisms beyond outcome optimization to support efficient and stable collaboration.

Keywords: agency; social coordination; joint commitment; human–AI collaboration

Introduction

A central challenge of cooperation is how agents coordinate their actions toward a joint goal when multiple, potentially conflicting goals are available in changing environments. Philosophical and evolutionary accounts argue that humans evolved species-unique forms of shared agency to solve such social coordination problems, enabling cooperation that is effective, flexible, and robust in ways that distinguish humans from other animals (Searle, 2003; Tomasello, 2019). A defining feature of human shared agency is its distinctive agentic structure. First, individuals form shared intentions (Bratman, 2013; Gilbert, 2017; Searle, 2010; Tomasello, 2022): jointly represented goals that are mutually recognized in the common ground. Coordination under shared agency is therefore not merely a coincidence of overlapping private goals, but a process of deciding together which joint goal to pursue among many possibilities, often supported by theory of mind (ToM) inference, communication and negotiation (Chater, 2023; Tomasello, 2025). Second, once established, a joint goal becomes the object of joint commitment (Gilbert, 2017), structuring partners' action regulation and expectations and sustaining coordination even as incentives change.

Importantly, shared agency emerges early in both evolution and ontogeny (Tomasello, 2022). Empirical research shows that 3-year-old children readily form shared goals, commit to them, and attempt to re-establish collaboration even when attractive alternatives become available or when they could achieve outcomes alone (e.g., Hamann et al., 2012). In contrast, nonhuman primates such as chimpanzees tend to treat partners instrumentally, disengaging once their individual incentives are met (e.g., Warneken et al., 2014). This contrast suggests that humans—already in early childhood—treat partners as intentional agents rather than as mere social tools, highlighting the uniquely social function of shared agency.

Contemporary artificial intelligence (AI), however, is designed along a fundamentally different trajectory from human shared agency. Despite rapid advances in performance and the emergence of increasingly powerful—and in some cases "alien"—forms of superhuman intelligence (e.g., Nobel Prize-winning breakthroughs in protein structure prediction; Jumper et al., 2021), current AI systems largely overlook the social intelligence that underpins human coordination (Chater, 2023; Duéñez-Guzmán et al., 2023; Tomasello, 2025). In most AI architectures, intelligence is framed in instrumental terms: systems are designed to optimize predefined objective functions, often via end-to-end learning approaches that abstract away from the social structure of interaction. Within this framework, no additional layer of "agency" is required between objective functions and rational action; artificial general intelligence (AGI)—or agency—is assumed to emerge as a byproduct of the efficient mapping of large-scale data and computation to optimal outcomes, largely independent of the agentic organization that constrains and structures human action.

As a result of the differences between humans and AI algorithms, most human–AI collaboration more closely resembles tool use (Farrell et al., 2025). Importantly, using AI tools (e.g., delivery robots and image generation tools (Saharia et al., 2022) effectively requires the human collaborator to adapt their behaviors to the tool, since the AI itself is not designed to interpret human intentions. More recent AI systems are designed to align with human preferences (Zhi-Xuan et al., 2025)—for example, through designing shared reward functions (see Du et al., 2023) or human-in-the-loop training (Christiano et al., 2017). These systems exhibit responsive behavior and support coordination via aligned utility structures or learned policies. However, alignment in either the asocial or social tool case does not amount to shared agency: such systems do not proactively interpret human intentions, negotiate or form shared goals, or maintain joint commitments. Instead, they treat others as instruments whose behavior can be predicted

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

or shaped to achieve system-defined objectives, rather than as partners in genuinely joint action.

This asymmetry of human shared agency vs. AI tool use raises a central set of questions for human–AI interaction. Can current AI agents effectively collaborate with humans to solve coordination problems that humans typically solve through shared agency? How efficient is human–AI collaboration, relative to human–human collaboration? To what extent is current human–AI interaction human-like and human-compatible?

These questions become especially sharp from a developmental perspective. While adults can typically flexibly switch between treating others as intentional agents and as tools, drawing on experience and strategic adaptation when interacting with instrumental systems (Dennett, 1989; Farrell et al., 2025), young children, in contrast, have limited experience treating social partners instrumentally. Children tend to view technology that is agent-like (e.g., robots, smart speakers, avatars) as capable of human-like collaborative behavior (Flanagan et al., 2023, 2024). As a result, young children may be especially challenged when successful collaboration requires adapting to AI systems that lack shared agency—not because children are poor collaborators, but because they are less willing or able to reinterpret a social partner as a tool. Child–AI interaction thus provides a critical test of whether an AI system genuinely supports shared agency or whether coordination success depends primarily on human compensation.

Here, we test different AI agents that collaborate with both adults and children using two mainstream AI frameworks: reinforcement learning (RL) and large language models (LLMs). Reinforcement learning provides a framework that integrates perception, actions, and utility functions, clearly embodying the instrumental view of intelligence. RL agents learn to maximize expected cumulative reward within a specified environment and can achieve highly optimal outcomes for the objectives they are trained on (Hubert et al., 2025; Jumper et al., 2021; Silver et al., 2017). Depending on their reward structure, RL may behave as asocial tools (optimizing individual rewards) or as social tools (optimizing joint rewards). Importantly, while joint rewards optimization can theoretically promote collective rational action (e.g., ant-like cooperation), such systems often struggle to maintain stable cooperation when multiple goals or equilibria exist (Chater et al., 2022; Tang et al., 2025; Zhao et al., 2022).

LLMs represent a distinct class of social tools. Although primarily trained on large-scale text corpora containing linguistic descriptions of human cooperation without real-time interaction or embodied coordination, LLMs can nevertheless generate behavior that appears socially sophisticated in text-based contexts. This stands in sharp contrast to human cognition, where shared agency is assumed to emerge prior to language and is grounded in pre-linguistic joint action (Tomasello, 2010). As a result, it remains unclear whether linguistic knowledge alone supports shared agency—particularly theory-of-mind–based coordination and joint commitment—or whether it can be translated into action in dynamic, non-linguistic environments. This concern is especially salient for visually grounded coordination tasks such as the ones we investigate here, where recent work suggests that multimodal LLMs fall short of human visual social reasoning abilities (e.g., Schulze Buschoff et al., 2025).

To address these questions, we designed a visual, non-linguistic sequential coordination game (Lewis, 1969/2008; Schelling, 1960). In this task, multiple equally valuable goals were distributed across a two-dimensional grid, and two players had to coordinate their actions to navigate to the same goal. Players can move one step to an adjacent cell when both players execute an action. A round is considered successful if both players reach the same goal. To probe coordination mechanisms beyond mere success outcomes, we designed several critical trials that focus on examining the consequences of shared agency (Fig. 1A). Specifically, at a critical shared intention moment where a joint goal is implicitly established through movement trajectories (i.e., both players clearly head toward the same goal), an alternative goal is introduced. This alternative goal is equally optimal from a utility perspective but suboptimal from a shared-intentionality perspective, in that it would require players to break the current joint commitment and establish a new one. From this moment onward, we measure two behavioral signatures of shared agency (Fig. 1B): (1) *Commitment*, measured by whether agents persist in reaching the previously established joint goal, and (2) *Signaling*, measured by whether agents choose actions that make their intended goal more transparent to their partner. We also measure *Efficiency* of such coordination, defined as the number of decisions made relative to the minimum required for success. Notably, during the game, the other player was introduced only as "another player," without specifying whether the partner was human or AI. We assessed participant perceptions of the AI player as more/less humanlike using a post-game questionnaire (we reported only descriptive statistics in Table 1 due to page limits).

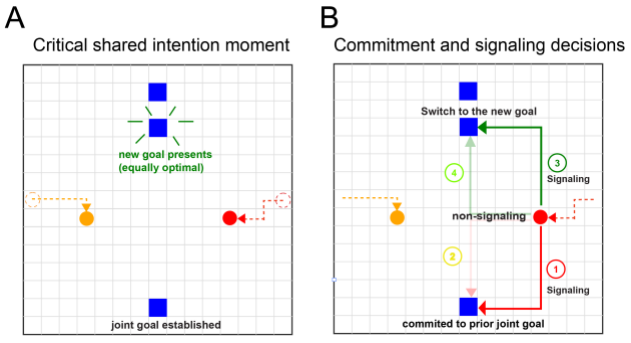

Figure 1: A. After both players' trajectories converge on the same goal, a new, equally optimal goal appears. Panel B. Players face choices of commitment and signaling. The four illustrated paths correspond to: (1) commit with signaling, (2) commit without signaling, (3) non-commit with signaling, and (4) non-commit without signaling.

Across two studies (adults in Experiment 1, and 5- to 7-year-olds in Experiment 2), we compare human–human collaboration with human–AI collaboration involving up to

four types of AI partners: Individual-RL agents, Joint-RL agents, standard LLM-based agents, and LLM-based agents augmented with ToM reasoning.

Experiment 1

Methods

Participants One hundred fifty adult participants ($N = 30$ per condition) were recruited via Prolific. Participants received approximately $3 for the 15-minute session, with the opportunity to earn a performance-based bonus of up to $1 (with each successful collaboration trial yielding $0.10). All procedures were approved by the Duke University Institutional Review Board and preregistered on AsPredicted here.

Materials and Design The experiment utilized a video-game-like 15 × 15 2D grid-world to display the sequential coordination game. Both players were represented as dots (one red and one orange), and goals were depicted as blue squares. Participants navigated the grid using the arrow keys on their own computer. Actions were executed simultaneously: the game state updated only after both players had made a movement choice, at which point both avatars moved by a single step to one of the four adjacent grid cells. Both players won the round only if they reached the same goal; if they reached different goals, the round was lost, and outcome feedback (win vs. loss) was provided at the end of every round. Participants were informed that earning more wins would increase their performance-based bonus.

The study employed a between-subjects design in which participants were randomly paired with a human or one of four AI partners.

(1) Individual-optimal RL agent. An agent optimized to maximize its own expected reward without representing or coordinating with the human partner (Baker et al., 2009; Cheng et al., 2023). Planning was formalized as a standard Markov decision process over the agent's own grid position. The agent computed an optimal action-value function:

$$Q(s,a) = \sum_{s'} T\,(s' \mid s,a)\left[R(s,a,s') + \gamma \max_{a'} Q(s',a')\right].$$

and selected actions via a softmax policy:

$$\pi(a \mid s) \propto \exp(\beta Q(s,a)).$$

where β controls choice stochasticity. β was fitted to 3 for all policies based on prior studies of similar 2D navigation tasks (Baker et al., 2009; Cheng et al., 2023).

(2) Joint-optimal RL agent. An agent optimized to maximize joint utility under the assumption of full cooperation, treating the human and AI as members of a single team (Shum et al., 2019). Planning was performed by a centralized joint planner that evaluated joint actions using a team-level action-value function:

$$Q^J(s,a_J) = \sum_{s'} T\,(s' \mid s,a_J)\left[R^J(s,a_J,s') + \gamma \max_{a'_J} Q^J(s',a'_J)\right].$$

where rewards were defined over joint outcomes. The agent then selected actions via a softmax policy over its marginal action values derived from the joint action-value function, executing only its own component of the selected joint action.

(3) Standard LLM-based agent. An LLM agent implemented using GPT-4.1-mini API (multimodal). At each time step, it received task instructions and the current game state (rendered visual map + symbolic coordinates + trajectory history) and output a single move from the set {up, down, left, right}.

(4) ToM–augmented LLM. A variant of the same model, with identical inputs and action space, but with one additional prompt: first, infer the partner's intended goal based on all available information, and then generate its action based on this inferred goal as well as the other available information.

Procedure The experiment consisted of four phases: 1) Solo Practice: Participants completed 3 trials with a single goal and 12 trials with two goals (and a possible new goal) to learn the basic movement mechanics. 2) Partner Matching: Participants were informed that they were being matched with "another player" online and experienced at least a 10-second waiting window to enhance plausibility. 3) Collaborative Practice: Participants completed 8 rounds with a partner; no new goals were introduced during this phase. 4) Critical Test Phase: Participants completed 12 test trials. Of these, 9 were critical test trials in which, once a joint goal could be inferred from the players' trajectories, a new, equally optimal third goal appeared. These critical trials were balanced such that the new goal was closer to the participant on 3 trials, closer to the partner on 3 trials, and equidistant on 3 trials. The remaining 3 trials did not include the appearance of a new goal, reducing participants' ability to infer the experimental manipulation.

Measures Collaboration success was coded as 1 if both players reached the same goal and 0 otherwise. Coordination processes included commitment, efficiency, and signaling. Commitment was coded as 1 if the human participant(s)' goal corresponding to the first detected shared intention matched the final goal reached by the dyad, and 0 otherwise. Efficiency captured relative path optimality following the appearance of a new goal and was computed by comparing the number of steps taken by the human participant from the onset of the new goal to the final goal against the shortest possible path for that segment: Efficiency = (1 − (actual steps − optimal steps) / actual steps) × 100. Signaling was coded as 1 if the human participant's first movement after the presentation of a new goal could be inferred as having a clear, goal-directed intent, and 0 otherwise.

Table 1: Human-Human vs. Human-AI Collaboration Success and Coordination Dynamics (Adults, $N = 30$ per group)

Measure	Human-Human	Human-Indiv. RL	Human-Joint RL	Human-LLM	Human-LLM-ToM
Overall success (%)	100.0 [100.0, 100.0]	96.9 [94.6, 99.1]	98.8 [97.5, 100.0]	79.3 [73.5, 85.1]*	88.9 [83.6, 94.2]*
Coordination Dynamics§					
Efficiency (%)‡	97.1 [95.3, 99.0]	91.7 [88.9, 94.5]	92.0 [88.2, 95.8]	80.2 [73.7, 86.7]*	84.0 [78.2, 89.8]*
Commitment (%)‡	84.4 [73.2, 95.6]†	56.7 [44.8, 68.5]*	71.1 [61.5, 80.8]	58.6 [47.3, 69.9]*	61.5 [48.1, 74.8]*
Signaling move (%)	51.3 [42.9, 59.7]	32.4 [26.7, 38.2]*	46.4 [39.7, 53.2]	17.2 [12.3, 22.0]*	43.0 [33.5, 52.5]*
Sign. & commit. (%)	53.1 [42.2, 64.1]	54.8 [45.7, 63.9]	60.6 [53.6, 67.6]†	32.8 [22.5, 43.1]	65.0 [52.5, 77.5]
Sign. & no commit. (%)	26.9 [13.8, 40.0]	6.0 [1.3, 10.7]*	7.5 [1.5, 13.4]*	3.0 [0.0, 6.5]*	20.0 [8.8, 31.3]
Perceived Agency *1 = definitely AI 5 = definitely human*	2.73 [2.36, 3.10]	1.90 [1.62, 2.18]*	2.43 [2.00, 2.87]	1.97 [1.61, 2.33]*	2.13 [1.72, 2.55]

Note. Values represent means (or percentages) computed from participant-level data. 95% confidence intervals are reported in brackets.
* Significantly lower than the Human-Human condition ($p < .05$).
† Significantly higher than the 50% chance baseline ($p < .05$).
‡ Mixed-effects modeling showed that commitment significantly predicted coordination efficiency across conditions ($\beta = 8.68$, $p < .001$).
§ The measures reported here focus on human behaviors.

Results

Collaboration success To test whether collaboration success differed by partner type, we fit a mixed-effects logistic regression with collaboration success as the binary outcome, partner type as a fixed effect, and participant ID as a random effect. The model revealed a significant effect of partner type on collaboration success ($\chi^2(4) = 62.59$, $p < .001$). Relative to the human–human condition, which achieved ceiling collaboration success, success rates declined substantially when participants were paired with both the standard LLM agent (OR = 45.62, $p < .001$) and the LLM-ToM agent (OR = 21.22, $p < .001$). In contrast, collaboration success did not differ reliably between the human–human condition and either RL partner (adult–Joint-RL: OR = 1.08, $p = .92$; adult–Individual-RL: OR = 1.19, $p = .83$). Comparisons between the two LLM-based agents indicated a marginal advantage for the LLM-ToM over the standard LLM (OR = 2.17, $p = .091$).

Commitment and coordination efficiency To test whether participants exhibit a tendency toward joint commitment under conflict—and whether such commitment supports more efficient coordination—we focused on successful collaboration trials in which the newly introduced goal was equally optimal at both the individual and joint levels.

We first examined whether commitment differed across partner types using a mixed-effects logistic regression with commitment as a binary outcome, partner type as a fixed effect, and participant ID as a random effect. This analysis revealed a significant effect of partner type on commitment to the initially inferred joint goal ($\chi^2(4) = 16.90$, $p = .002$). Human-human pairs exhibited the highest level of commitment (84.4%), reliably persisting toward their shared plan despite the availability of an alternative. Direct tests against the 50% chance baseline using estimated marginal means with Bonferroni adjustment confirmed that only the human–human condition showed a significant commitment bias relative to chance ($\hat{p} = 0.859$, SE = 0.041, $z = 3.89$, $p = .0005$), whereas commitment in all human–AI conditions did not differ reliably from chance (all ps > .45). Post-hoc comparisons further indicated reduced commitment in human–AI interactions, particularly when participants were paired with the Individual-RL agent (OR = 4.63, $p = .002$), the standard LLM agent (OR = 4.28, $p = .005$), and the LLM-ToM agent (OR = 3.36, $p = .036$).

We next assessed whether coordination efficiency varied as a function of partner type and commitment using a linear mixed-effects model predicting efficiency from partner type and commitment, with participant ID treated as a random effect. This analysis revealed a robust effect of partner type on coordination efficiency, with efficiency significantly lower in human–AI pairs than in human–human pairs. In particular, efficiency was substantially reduced in the standard LLM condition ($\beta = -15.27$, $p < .001$) and in the LLM-ToM condition ($\beta = -11.89$, $p < .001$), whereas differences between the human–human condition and either RL partner were not reliable. Crucially, commitment emerged as a strong independent predictor of coordination efficiency ($\beta = 8.68$, $p < .001$), indicating that persisting toward an established joint goal substantially improved efficiency across partner types.

Signaling To examine whether signaling behavior differed across partner types, we fit a mixed-effects logistic regression with signaling moves as a binary outcome, partner type as a fixed effect, and participant as a random intercept. The model revealed a significant main effect of partner type ($\chi^2(4) = 37.46$, $p < .001$). Relative to the human–human condition, participants signaled significantly less when interacting with the Individual-RL agent ($\beta = -0.79$, $p < .001$), the standard LLM agent ($\beta = -1.65$, $p < .001$), and the LLM-ToM agent ($\beta = -0.48$, $p = .025$). In contrast, signaling did not differ reliably between human–human pairs and human-Joint-RL pairs ($\beta = -0.17$, $p = .40$). Post-hoc pairwise comparisons using Tukey adjustment further showed that the Joint-RL agent elicited significantly more signaling than the Individual-RL agent (OR = 1.86, $p = .022$), and that participants signaled more when interacting with the LLM-

ToM agent than with the standard LLM (OR = 3.22, p = .002).

To further examine whether signaling depended on participants' commitment state, we fit a mixed-effects logistic regression including commitment, partner type, and their interaction, with participant as a random intercept. This analysis revealed a strong main effect of commitment ($\chi^2(1) = 115.47$, $p < .001$), as well as a significant partnerType × commitment interaction ($\chi^2(4) = 25.75$, $p < .001$), indicating that differences in signaling across partner types depended on whether participants were committed. Post-hoc Tukey-adjusted comparisons showed that when participants were uncommitted (i.e., switched to a new goal), signaling was highest in the human–human condition and significantly reduced for the Individual-RL agent, the Joint-RL agent, and the standard LLM (all ps < .01), but not for the LLM-ToM agent (p = .22). When participants were committed, signaling rates increased substantially across all partner types, and differences between partner types were greatly attenuated. Notably, only the Joint-RL agent elicited signaling reliably above the chance level of 50% (M = .61, 95% CI [.54, .68], Bonferroni-adjusted p = .021).

Discussion

These findings suggest that adults can collaborate successfully with tool-like partners that implement optimal decision policies, even if their actions are driven by self-interest (i.e., individual RL). In contrast, coordination with LLM-based agents is substantially less effective in this visually grounded task. Although ToM-like cues provided a marginal improvement over a standard LLM, they did not approach human-level or optimal RL-level performance, consistent with prior findings that LLMs struggle with visual cognition (Schulze Buschoff et al., 2025). Beyond overall success, human–AI interactions were marked by reduced joint commitment, which reliably predicted lower coordination efficiency across partner types.

Signaling behavior revealed a further distinction in how adults adapted to different partners. Participants signaled most reliably when interacting with human partners and with the Joint-RL agent, and were especially likely to signal when committed in the Joint-RL condition. This pattern suggests an uncanny-valley–like effect (Mori, 1970): the Joint-RL agent is perceived as somewhat human-like due to its competent and responsive behavior, yet its coordination with humans remains less smooth and its commitment uncertain, prompting more proactive signaling to stabilize joint action.

Experiment 2

The adults in Experiment 1 may have been using a variety of sophisticated strategies for coordination, actively compensating for the shortcomings of their partner in complex ways. Although children are already skillful at coordinating with others from an early age—first with adults and later with both adults and peers after age 3 (Tomasello, 2019)—younger children may be more likely to assume by default that interactive partners are well-intentioned and capable agents (Flanagan et al., 2023; 2024). As a result, they may adopt similar coordination strategies across partner types, which could lead to poorer coordination when a partner exhibits low levels of agency-like behavior. As in the adult study, only the RL agents—not the LLM-based agents—achieved high levels of collaboration success, demonstrating that they were capable tools for the current task. In Experiment 2, therefore, we focus on investigating how 5- to 7-year-old children interact with Individual-RL and Joint-RL partners, preserving the key theoretical contrast between asocial and social tools.

Methods

Participants The sample consisted of 30 5– to 7-year-old children (M_{age} = 6.68, SD_{age} = 0.99, 53.3% females). There were 15 children in each condition. Children were recruited from a lab database or a local museum in a small city in the Northeastern United States. All respondents reported their child's ethnicity, with 63.33% identifying as White, 30.00% as multiracial, 6.67% as Asian, and 6.67% as Hispanic or Latino. Most caregivers reported an annual household income above $100,000 USD (96.43%, n = 27) and held a graduate or professional degree (93.33%, n = 28).

Materials, Design and Procedure The study was conducted either in person at a local museum or remotely via Zoom. Children were randomly assigned to one of two conditions (Individual-RL or Joint-RL). At the beginning of each phase, the experimenter introduced the game rules and demonstrated the task to ensure comprehension. Following each instruction phase, children were asked three comprehension questions by the experimenter to confirm their understanding before proceeding (all children passed). The game procedure closely mirrored the adult version but used fewer trials to reduce session length and maintain engagement. Across the four phases, children completed 2, 8, 4, and 8 trials, respectively (corresponding to Solo Practice with one goal; Solo Practice with two goals; Collaborative Practice; and the Critical Test Phase).

Results

Collaboration success We examined whether children's collaboration success differed as a function of AI partner type and whether these differences varied across age. We fit a mixed-effects logistic regression with collaboration success as the binary outcome, partner type, age, and trial index as fixed effects, and participant ID as a random intercept. The interaction between partner type and age showed a marginal trend ($\chi^2(1) = 2.73$, p = .098), suggesting that age-related changes in collaboration success differed between the Individual-RL and Joint-RL conditions (Figure 2). To interpret the interaction, we examined model-estimated collaboration success rates at specific ages. At age 5, children were more likely to succeed when paired with the Joint-RL agent (M = 0.93, 95% CI [0.67, 0.99]) than with the Individual-RL agent (M = 0.67, 95% CI [0.38, 0.87]), although this difference did not reach statistical significance

(OR = 6.25, p = .10). By age 7, collaboration success converged across partner types (OR = 1.09, p = .89), with both conditions showing high success rates.

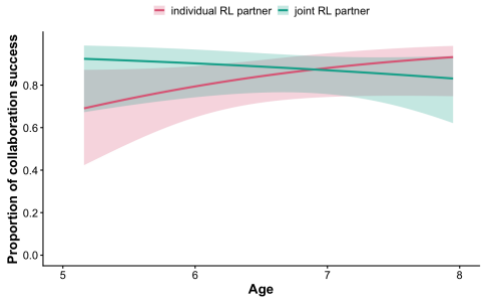

Figure 2: Predicted probability of collaboration success across age. Solid lines show model-predicted probabilities from a mixed-effects logistic regression, and shaded bands indicate 95% confidence intervals.

Commitment and coordination efficiency To test whether commitment also contributes to children's coordination efficiency, we fit a linear mixed-effects model predicting coordination efficiency from commitment, partner type, age, and trial order, with each participant treated as a random effect. Commitment showed a marginal association with coordination efficiency ($\chi^2(1) = 3.28$, $p = .070$), such that trials in which children persisted with the initially inferred joint goal tended to be more efficient than trials in which they switched to the alternative goal.

Signaling We fit a mixed-effects logistic regression with signaling moves as the binary outcome, partner type, age, trial index, and commitment as fixed effects, and participant as a random effect. This analysis revealed significant main effects of partner type ($\chi^2(1) = 9.51$, $p = .002$) and commitment ($\chi^2(1) = 30.79$, $p < .001$), but no reliable effects of age ($\chi^2(1) = 0.28$, $p = .59$) or trial index ($\chi^2(1) = 0.03$, $p = .86$). Tukey-adjusted pairwise comparisons showed that signaling was significantly more frequent in the Joint-RL condition than in the Individual-RL condition (OR = 4.17, $p = .043$). Moreover, children were more likely to signal on trials in which they ultimately maintained their prior commitment than on trials in which they switched to the alternative goal (OR = 33.33, $p < .001$).

Discussion

These preliminary findings suggest a developmental shift in how children adapt to different types of AI partners, with younger children tending to collaborate more successfully with a Joint-RL agent and older children achieving high success with both asocial and social tools. Commitment showed a modest association with coordination efficiency, such that children tended to coordinate more efficiently when they persisted with an initially inferred joint goal, indicating an emerging role of commitment for efficient coordination.

Surprisingly, although younger children struggled to adapt to asocial tools—a task that appears relatively easy for adults—they nevertheless displayed seemingly sophisticated signaling strategies. This pattern suggests that children's ability to engage in collaboration as equal partners through communication and coordination may be more foundational, whereas treating an interaction partner as a mere tool may require more advanced strategies that emerge with age.

General Discussion

The present studies examined whether current AI systems support human-like coordination by testing their ability to collaborate with adults and children in a non-linguistic sequential coordination game targeting core components of shared agency. Across two experiments, we observed a clear dissociation between coordination success and the agentic structure underlying coordination. Although adults often achieved high success rates with AI partners—particularly optimal RL agents—only human–human dyads consistently exhibited behavioral signatures of shared agency: they committed to a joint goal, which made their collaboration most efficient. These findings support the claim that outcome optimization alone is insufficient to characterize shared agency, and that much of human–AI collaboration more closely resembles tool use than genuine joint action.

Crucially, this dissociation was evident developmentally. In Experiment 2, 5-year-old children benefited from interaction with Joint-RL agents, whereas by age 7 children performed comparably with Joint- and Individual-RL partners. These age-related differences suggest that children initially expect collaborators to engage in human-like shared agency by default, but with age become increasingly able to adopt instrumental strategies that succeed even when partners act solely on individual utilities.

Patterns of signaling further revealed how humans adapt to different partners. Signaling was highest in human–human interaction but was also elevated for Joint-RL relative to other AI tools. This pattern suggests that signaling is sensitive to expectations about whether one's actions will be interpreted and reciprocated, rather than being a fixed feature of cooperation. In particular, elevated signaling toward Joint-RL may reflect compensatory communication in response to partial agency cues: when a partner appears competent and responsive, yet lacks fully reliable commitment or smooth coordination, individuals may increase explicit signaling to stabilize joint action.

More broadly, these findings point to the value of moving beyond outcome optimization toward reverse engineering embodied shared agency. Evaluating and modeling human agentic structures grounded in evolution and ontogeny may provide insights into the development of human-compatible AI and AGI (Tomasello, 2025). Developing AI systems that function not merely as tools, but as autonomous collaborative partners, will likely require mechanisms that support shared goals, manage tensions between individual and collective objectives, and make coordination dynamics transparent and intelligible to human partners. Advancing such mechanisms may enhance efficiency while also strengthening trust, interpretability, and alignment in human–AI interaction, enabling AI systems to participate in collaboration in ways that are both effective and socially meaningful.

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*(3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Bratman, M. E. (2013). *Shared agency: A planning theory of acting together*. Oxford University Press.

Chater, N. (2023). How could we make a social robot? A virtual bargaining approach. *Philosophical Transactions of the Royal Society A, 381*(2251), 20220040. https://doi.org/10.1098/rsta.2022.0040

Chater, N., Zeitoun, H., & Melkonyan, T. (2022). The paradox of social interaction: Shared intentionality, we-reasoning, and virtual bargaining. *Psychological Review, 129*(3), 415–435. https://doi.org/10.1037/rev0000343

Cheng, S., Zhao, M., Tang, N., Zhao, Y., Zhou, J., Shen, M., & Gao, T. (2023). Intention beyond desire: Spontaneous intentional commitment regulates conflicting desires. *Cognition, 238*, 105513. https://doi.org/10.1016/j.cognition.2023.105513

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems, 30*, 4299–4307.

Dennett, D. C. (1989). *The intentional stance*. MIT Press.

Du, Y., Leibo, J. Z., Islam, U., Willis, R., & Sunehag, P. (2023). A review of cooperation in multi-agent learning. *arXiv preprint arXiv:2312.05162*. https://doi.org/10.48550/arXiv.2312.05162

Duéñez-Guzmán, E. A., Sadedin, S., Wang, J. X., McKee, K. R., & Leibo, J. Z. (2023). A social path to human-like artificial intelligence. *Nature Machine Intelligence*, 5(11), 1181-1188. https://doi.org/10.1038/s42256-023-00770-9

Farrell, H., Gopnik, A., Shalizi, C., & Evans, J. (2025). Large AI models are cultural and social technologies. *Science, 387*(6739), 1153–1156. https://doi.org/10.1126/science.adt9819

Flanagan, T., Georgiou, N. C., Scassellati, B., & Kushnir, T. (2024). School-age children are more skeptical of inaccurate robots than adults. *Cognition*, 249, 105814. https://doi.org/10.1016/j.cognition.2024.105814

Flanagan, T., Wong, G., & Kushnir, T. (2023). The minds of machines: Children's beliefs about the experiences, thoughts, and morals of familiar interactive technologies. *Developmental Psychology*, 59(6), 1017. https://doi.org/10.1037/dev0001524

Gilbert, M. (2017). Joint commitment. In *The Routledge handbook of collective intentionality* (pp. 130–139). Routledge. https://doi.org/10.4324/9781315768571-13

Hamann, K., Warneken, F., & Tomasello, M. (2012). Children's developing commitments to joint goals. *Child Development, 83*(1), 137–145. https://doi.org/10.1111/j.1467-8624.2011.01695.x

Hubert, T., Mehta, R., Sartran, L., Horváth, M. Z., Žužić, G., Wieser, E., ... & Silver, D. (2025). Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature, 651(8106), 607–613*. https://doi.org/10.1038/s41586-025-09833-y

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Lewis, D. (2008). *Convention: A philosophical study*. John Wiley & Sons. (Original work published 1969).

Mori, M. (1970). *The Uncanny Valley.* Energy, 7(4), 33–35.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., … Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems, 35*, 36479–36494. https://doi.org/10.52202/068431-2643

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

Schulze Buschoff, L. M., Akata, E., Bethge, M., & Schulz, E. (2025). Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 7(1), 96-106. https://doi.org/10.1038/s42256-024-00963-y

Searle, J. R. (2003). *Rationality in action*. MIT Press.

Searle, J. R. (2010). *Making the social world: The structure of human civilization*. Oxford University Press.

Shum, M., Kleiman-Weiner, M., Littman, M. L., & Tenenbaum, J. B. (2019). Theory of minds: Understanding behavior in groups through inverse planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 6163–6170. https://doi.org/10.1609/aaai.v33i01.33016163

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354-359. https://doi.org/10.1038/nature24270

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. https://doi.org/10.1098/rspb.2024.3070

Tomasello, M. (2010). *Origins of human communication*. MIT Press.

Tomasello, M. (2019). *Becoming human: A theory of ontogeny*. Harvard University Press.

Tomasello, M. (2022). *The evolution of agency: Behavioral organization from lizards to humans*. MIT Press.

Tomasello, M. (2025). How to make artificial agents more like natural agents. *Trends in Cognitive Sciences, 29*(9), 783–786. https://doi.org/10.1016/j.tics.2025.07.004

Warneken, F., Steinwender, J., Hamann, K., & Tomasello, M. (2014). Young children's planning in a collaborative problem-solving task. *Cognitive Development, 31*, 48–58. https://doi.org/10.1016/j.cogdev.2014.02.003

Zhao, M., Tang, N., Dahmani, A. L., Zhu, Y., Rossano, F., & Gao, T. (2022). Sharing rewards undermines coordinated hunting. *Journal of Computational Biology*, 29(9), 1022–1030. https://doi.org/10.1089/cmb.2021.0549

Zhi-Xuan, T., Carroll, M., Franklin, M., & Ashton, H. (2025). Beyond Preferences in AI Alignment. *Philosophical Studies*, *182*(7), 1813-1863. https://doi.org/10.1007/s11098-024-02249-w