Intentional commitment as a spontaneous presentation of self

Shaozhe Cheng[1], Jingyin Zhu[1], Jifan Zhou*[1], Mowei Shen*[1], Tao Gao*[2,3,4]

[1]Department of Psychology and Behavioral Sciences, Zhejiang University

[2]Department of Communication, UCLA

[3]Department of Statistics, UCLA

[4]Department of Psychology, UCLA

Author Note

The authors have no conflicts of interest to disclose. Some of the data in this work were previously presented at the Workshop on Social Intelligence in Humans and Robots during the annual meeting of the Robotics: Science and Systems 2022. The design and analysis plans for Experiment 3 were preregistered at osf.io/gzvf3.

This work was supported by UCLA startup funding to Tao Gao, and grants from National Natural Science Foundation of China (32071044 to Mowei Shen, 32371088 and 6233700 to Jifan Zhou). The authors also wish to acknowledge anonymous reviewers for helpful comments on previous drafts.

*Correspondence concerning this article should be addressed to be addressed to

Tao Gao, Department of Communication, UCLA, Los Angeles, CA 90095. Email: taogao@ucla.edu

Jifan Zhou or Mowei Shen, Department of Psychology and Behavioral Sciences, Zhejiang University, Zijingang Campus, 866 Yuhangtang Road, Hangzhou, ZJ 310058, P.R. China. Email: jifanzhou@zju.edu.cn or mwshen@zju.edu.cn

Abstract

Commitment is a defining feature of human rationality. This study explores a social origin of spontaneous intentional commitment, assuming commitment in individual decision-making arises from an internalized self-presentation, transferring the audience of commitment from a real partner to an inner eye perspective. To test this "social inner eye" hypothesis, we exposed participants to different social contexts while maintaining the individual nature of the task. Across three experiments, we found that (a) individuals consistently showed stronger commitment when acting in front of others, (b) different social contexts had different impacts on the process of commitment formation, with the mere outside observer accelerating commitment, while a parallel player delays it, (c) participants spontaneously coordinated their intentions to avoid conflicts when playing with another parallel player, despite no coordination was required. Taken together, we demonstrated how social context influences the strength, content, and timing of individual commitment. These findings align with the perspective that individual commitment has a social origin. They also contribute to an understanding of why commitment is universally valued across cultures and is seen as a virtue rather than a weakness in human decision-making.

Keywords: commitment, intention, self-presentation, theory of mind, social origin

Statement of limitations

Our study faced certain limitations that could serve as a foundation for future research. First, our methodology incorporated a video game-like task within a controlled, lab-based environment. While this approach ensures experimental rigor and control over extraneous variables, future studies could explore how behaviors observed in this simulated environment generalize to real-life tasks and settings. For example, in our experiments, we represented two conflicting desires as visual shapes with clearly defined and identical rewards. However, in the real world, personal goals often have unique attributes, such as working in the lab versus playing video games, the rewards of which are not easily quantifiable. The applicability of our findings to such real-world conflicting goals warrants further investigation. In addition, the participants in our studies were adults in China. Although we discussed the "social inner eye" as a universal cognitive process, it is possible that different social environments and cultures (e.g., Eastern versus Western) may have different socialization pathways for shaping individual commitment, which could also lead to different developmental trajectories (e.g., Chernyak et al., 2019). Thus, it is important for future studies to explore the developmental and cultural aspects of the "social inner eye" hypothesis.

Introduction

Commitment presents a paradox in human life. Within the traditional contexts of behavioral economics and organizational psychology, commitment is often studied as an irrational bias, termed the escalation of commitment or sunk cost fallacy, which causes individuals to persist with decisions despite suboptimal outcomes (Staw, 1976; Arkes & Ayton, 1999). On the other hand, in everyday life, commitment is widely regarded as a virtue essential for achieving challenging goals, at both the individual and collective levels. The appreciation of commitment is well demonstrated by Winston Churchill's inspiring speech, in which he urged perseverance in the face of difficulty: "...never give in. Never, never, never, never—in nothing, great or small, large or petty…" Echoing this appreciation for commitment in real life, philosophers have long emphasized it as a defining characteristic of human rationality, rather than viewing it as a weakness of human nature. It has been argued that while animal rationality predominantly focuses on finding the optimal *means* to satisfy certain desires, the unique aspect of human rationality lies in their ability to commit to an *end* in the face of complex, even conflicting desires (Bratman, 1987; Searle, 2003). Such ability is essential in human life; consider, for instance, the scenario of planning a summer vacation to several enticing destinations. Faced with such choices, an individual needs to decide on and commit to a single location to formulate a coherent plan. Commitment, as a resolution of conflicting desires, enables individuals to act coherently following a chosen path and to resist the temptation of competing desires.

Previous studies, using introspection and subjective self-reports, have found that leveraging commitment as a high-level strategy can facilitate goal attainment (Gollwitzer & Brandstätter, 1997; Gollwitzer, 1999; Gollwitzer & Sheeran, 2006; Ajzen et al., 2009; Nenkov & Gollwitzer, 2012; Ajzen & Kruglanski, 2019). For example, when individuals are encouraged or explicitly instructed to make a commitment—such as signing a commitment statement or setting a New Year's resolution—they are more likely to achieve their intended future goals, like quitting smoking or losing weight.

Recent studies have started to explore the phenomenon of spontaneous commitment in the context of rational decision-making, where both the goal value and the action cost are clearly defined. One study found that when faced with two valuable moral goals, both adults and children (aged 4-6 years) tend to stick with their chosen goal even if it becomes costly later on (Chu & Schulz, 2022). Another study examines how commitment regulates conflicting desires in sequential decision-making tasks (Cheng et al., 2023). This study draws inspiration from the Buridan's ass paradox, a thought experiment hypothesis that an ass, lacking intention, might starve when faced with two equally desirable piles of hay due to indecision. The findings indicate that humans resolve conflicting desires by committing to an intention, making their actions qualitatively different from those of a purely desire-driven agent, which acts solely to maximize expected utilities. A subsequent study further demonstrated that 6-year-olds, but not 5-year-olds, spontaneously commit to a future goal, even if new opportunities emerge that make their initial choice less optimal. Moreover, they found a positive correlation between children's intentional commitment and their executive control (Zhai et al., 2023).

Collectively, these recent studies suggest that humans spontaneously use their executive control to establish an intentional commitment, guiding their actions towards achieving their intended goal—even when such a commitment is not explicitly mandated in a task. These studies also support the argument that human intentional actions are beyond a pure reward maximization model (e.g., reinforcement learning), highlighting intention as a critical intermediate mental state for understanding almost every aspect of human decision-making, ranging from how we represent the external world and establish internal values to the final decisions we make (Molinaro & Collins, 2023). The goal of this study is to explore the origins of intentional commitment as well as the pattern of human behavior both antecedent to, and subsequent to, the establishment of commitment.

Social inner eye: commitment as a spontaneous presentation of self

It has been suggested that one important function of commitment is that it makes one's future actions more predictable, thus facilitating social coordination (Bratman, 1987; Michael & Pacherie, 2015). Findings show that, unlike chimpanzees, children as young as three can form joint commitments to regulate their own behavior (Graefenhain et al., 2009; Hamann et al., 2012; Duguid et al., 2014), leading to an argument that social self-regulation is essential for understanding unique human cooperation (Tomasello, 2022). Here, we explore the possibility that the constraints of commitment transcend the original context of social cooperation and manifest spontaneously even in individual tasks, making the social context an important contributor to individual commitment. This hypothesis builds on the seminal idea that many features of individual actions stem from the internalization of cognitive capacities that originally developed for social interactions (Vygotsky, 1930). Mead (1934) speculated that humans learn to perceive themselves through the eyes of others. He highlighted the human ability to imagine themselves in the role of the other and to take the other’s perspective on themselves. Importantly, this viewpoint can transition from a specific social other to a broader “generalized other”, which in essence represents the perspective of society itself. In this sense, humans develop self-consciousness and can view themselves from a “bird’s eye” (Tomasello et al., 2005) or an “inner eye” perspective (Humphery, 2002). Consequently, even when individuals act alone, they may still spontaneously reflect on their own actions from a third-party objective perspective.

This "social inner eye" phenomenon has long been identified as the phenomenon of imagined audiences by sociologists and developmental psychologists, where people imagine and believe that others are watching them, even when no one is around (Goffman, 1959;

Elkind, 1967). Critically, modern theories of metacognition suggest that individuals could exploit such an imaginary audience for self-monitoring and self-reflection (Bandura, 1989; Flavell, 1999; Rochat, 2009; Frith & Heyes, 2012; Shea et al., 2014). The capacity of social imagination is also crucial for moral judgement. Recent studies show that by age six, children start to evaluate those who act with the imaginative others in mind as "nicer," even when they are merely pursuing their own individual goals (Zhao et al., 2021; Kushnir, 2022).

Building on these views, we propose that commitment in individual decision-making arises from an internalization of commitment originally developed in social contexts, transferring the audience of commitment from a real partner to an inner eye viewed from a third-party perspective. This "social inner eye" (SIE) hypothesis predicts that when participants are placed in a social context with others present—highlighting the public nature of their actions—the commitment would be further enhanced. In addition, the presence of social others may trigger additional intention coordination processes where humans spontaneously read others' intentional commitments and adjust their own accordingly, even in tasks that are inherently individual and require no coordination.

In three experiments, we examined whether the strength and dynamic process of commitment would be affected by the presence of a second person (Figure 2), which was introduced as having a real person sit next to the participant (Experiment 1) or having two participants play their own individual task in parallel on the same display (Experiment 2 and Experiment 3). Moreover, in the two-player experiments, we further explore whether participants spontaneously coordinate their actions by testing how participants' own commitment was influenced by the manifested intentions of another agent.

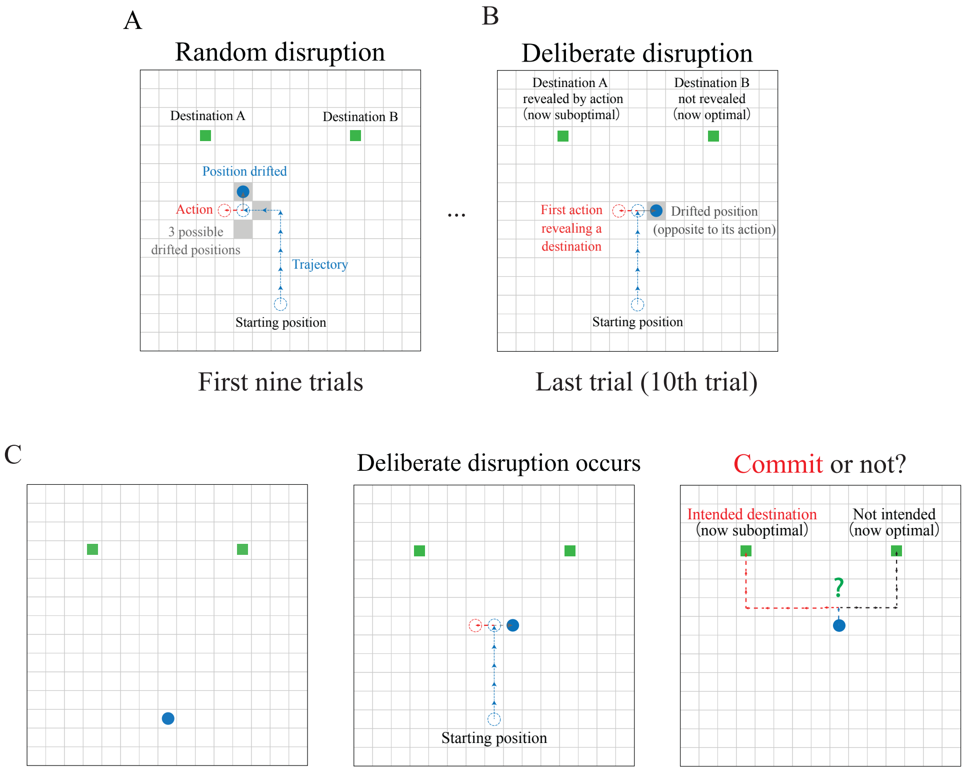
General Method

We adopted the "goal perseverance" paradigm to measure intentional commitment (Cheng et al., 2023). Within this paradigm, commitment was measured as an implicit bias; it was neither mentioned nor required in the task, and could even hurt the main task performance. It was defined as the persistence in pursuing the initially intended goal, even when unexpected disruptions rendered it suboptimal. In the experiment, participants were asked to navigate an agent towards one of two equally appealing destinations that were distantly positioned on a 2D grid world. Their task was to reach the destination as quickly as possible while taking the fewest steps. Participants were told that they could earn an additional bonus if they finished the task efficiently following the instructions.

During navigation, unexpected disruptions could occur that negated the agent's planned action, causing the agent to drift to a nearby unintended location (Figure 1A). At a certain moment, a deliberately engineered disruption inadvertently moved the agent closer to the unchosen destination (Figure 1B). "Goal perseverance" was measured by the agent's decision to continue towards its originally intended destination, despite the unintended one now being closer (Figure 1C). To prevent arousing participants' suspicion, the deliberate disruption was introduced only once in the last trial as the "critical trial" (Mack & Rock, 1998) of the experiment. This disruption was disguised by the first nine trials of random disruptions, which served to establish the impression that all disruptions were "random," according to the instructions given to the participants. Additionally, the first nine trials were used to establish a baseline of the rationality of human actions by calculating the average percentage of trials in which participants reached a suboptimal goal in random disruption trials. Data from all ten trials were also used to reveal the temporal dynamics of intentional commitment by inferring intentions from actions from a theory of mind perspective.

Figure 1

Measure commitment as "goal perseverance"

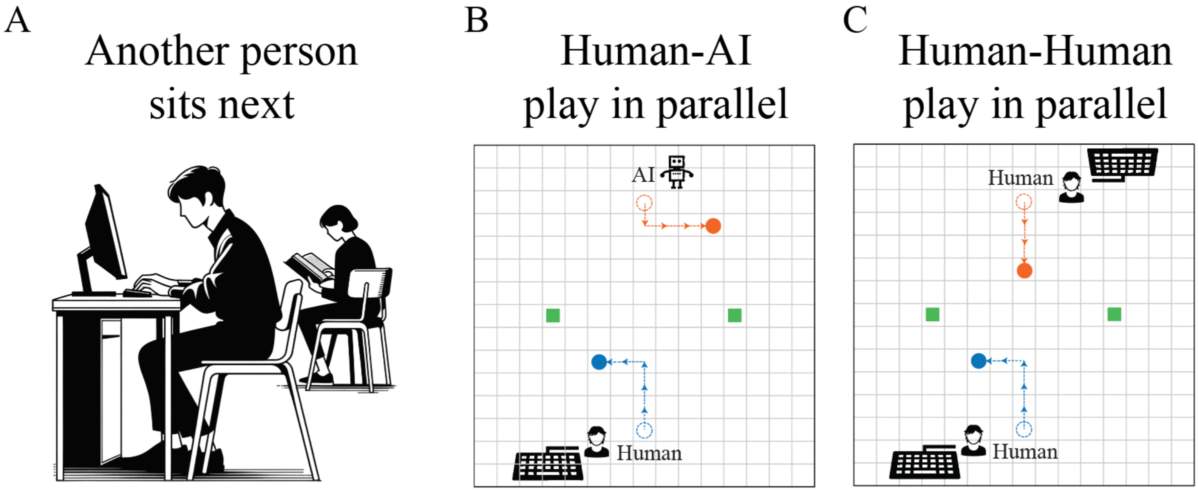


Note. Panel A. In the first nine trials, both the time step and the direction of the disruptions were randomly sampled. Panel B: Design of deliberate disruptions. In the last trial, both the time step and the direction of the disruptions were deliberately designed to push the agent away from the destination the moment it was revealed. Panel C: This is a sampled critical trial with deliberate disruption. Commitment is measured by whether the agent would still pursue the originally intended but now sub-optimal destination after the deliberate disruption. All dashed lines are for illustration; they are not visible in the real experiment.

To measure the impact of social context on individual commitment, we compared our results with those from Cheng et al.'s (2023) Experiment 1, whose main task was identical to ours, except that their participants completed the task in a private room, where their real-time actions were not observed by others. Our analysis went beyond mere examination of the increase or decrease in overall task performance, which could be attributed to simple arousal or engagement due to social presence. We aimed to test whether the social context could specifically increase "goal perseverance" in the last critical trial, thereby dissociating it from the effects on task performance in the first nine trials. Moreover, as our results will show, different social contexts could even exert opposite impacts on different stages of commitment. These effects were entirely independent of the main task performance, which remained unaffected by the social context throughout the experiments.

Figure 2

Manipulations of public context in an individual task



Note. Panel A. In Experiment 1, a second person will be present in the same room, sitting diagonally behind the participant, reading a book. Panel B. In Experiment 2, a second player will be present in the game, playing the same individual task in parallel with participants. Participants were told this second agent was controlled by another human, yet it was actually controlled by a softmax MDP-AI agent. Panel C. In Experiment 3, the second player was actually controlled by another human participant.

Transparency and Openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. Data is available at https://osf.io/rca3j/?view_only=60c2a2d90720457ea1689e9d3c81b208. The design and analysis plans for Experiment 3 were preregistered at https://osf.io/gzvf3/?view_only=9c51337dec84440dad5d199679bc4111

Experiment 1: Mere presence of a second person

In this experiment, participants completed the navigation task while a second person was present in the room (Figure 2A). We intentionally introduced only a weak social context, wherein the second person was instructed to sit quietly and read a book, without interacting with the participants or directly observing them. To avoid arousing the participants' suspicions, this individual was introduced as an experimental assistant who would remain in the room to assist with any computer-related issues. Both the participants and this second person were unaware of the experiment's true purpose.

Method

Participants

The sample size was set at 50 to match that of Cheng et al.'s (2023) Experiment 1. This sample size provided 80% power to detect an odds ratio = 1.77 or greater in a logistic regression test with a 5% false-positive rate (by G*Power 3.1). A total of 50 adult participants (M_{age} = 22.2, range = 18-25) were recruited from the participant pool at XXX University for the experiment. They were paid 10 RMB for participation, and an additional 5 RMB for completing the task following instructions.

This experiment, as well as the subsequent ones, was pre-reviewed and approved by the Institutional Review Board of the Department of Psychology at XXX University. All participants in this study provided informed consent prior to participating in the experiments. No participants were excluded from the analysis.

Design and Procedure

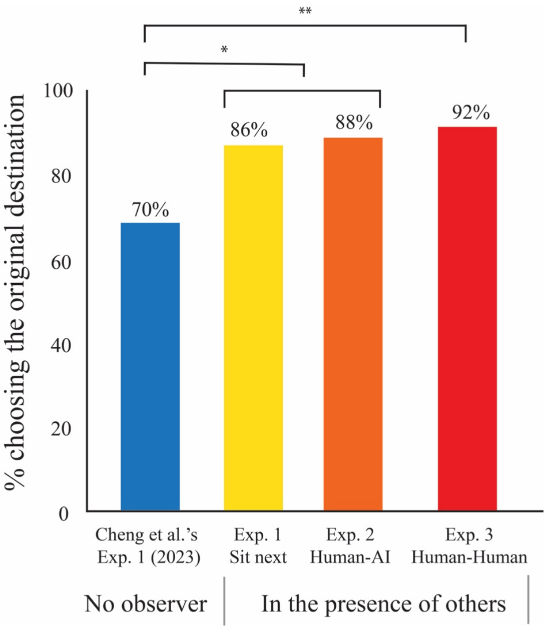
There were 10 trials in total. Each trial consisted of a 15×15 grid map presenting an agent (RGB: 50, 50, 255) and two destinations (RGB: 255, 50, 50). The destinations were placed such that the Manhattan distances from the agent's starting position to each destination were equal. Participants were explicitly informed that the environment was not deterministic: at each step, there was a 10% probability that the agent's action could be interrupted by a random drift pushing it to a nearby cell. In the initial nine trials, as per the instructions given to the participants, disruptions occurred randomly, resulting in approximately one drift per trajectory. The disruption in the final trial was not random, but only triggered when the agent first indicated its destination by moving towards one destination while away from the other.

Results

We first examined the rationality baseline. In the first nine trials with random disruptions, participants reached a suboptimal goal in 3.8% of trials (95% CI: [0.021, 0.054]). This ratio did not statistically differ from Cheng et al.'s (2023) no-observer experiment (3.8% vs 4.2%, $t(98) = 0.32$, $p = .7455$, Cohen's $d = 0.07$, $BF_{10} = 0.22$). However, in the last trial with deliberate disruption, 86% of participants reached the originally-intended but later-suboptimal goal (Figure 3), which was significantly higher than in Cheng et al.'s (2023) no-observer experiment (86% vs. 70%, Fisher's exact test, odds ratio = 2.63, one tailed $p = .0448$, Cramer's $\varphi = 0.19$; $BF_{10} = 1.88$). Together, these results suggest that adding a potential observer had no general impact on the optimality of human actions in random disruption trials, but had an effect of reducing performance in the deliberate disruption trial due to a higher percentage of "goal perseverance". Considering that the social context introduced here was only minimal, we aimed to introduce different social contexts in subsequent experiments to see whether we can replicate and extend this finding.

Figure 3

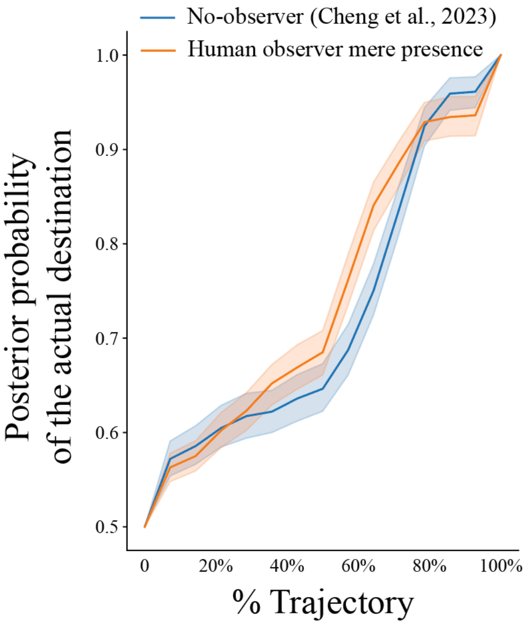
Enhanced "goal perseverance" in the presence of others



Note. Percentage of participants who reached the originally intended destination in the last deliberate disruption trial. *p < .05, **p < .01

Figure 4

Human intentions in the eyes of a BToM observer



Note. The posterior of the BToM inference for the agent's actual destination as a function of agent's steps over time. The error shadows reflect 95% confidence intervals.

Human intention in the eyes of an observer

In addition, to explore how the presence of others impacted the dynamic process of intention formation, we applied a Bayesian Theory of Mind (BToM) model to infer human's intention from their actions using data from all trials (Baker et al., 2009):

$$P(\text{ Intention} \mid \text{Action}_{1:T}, \text{Environment}) \propto \prod_{t=1}^{T} P(\text{Action}_t \mid \text{Intention, Environment}) P(\text{Intention} \mid \text{Environment})$$

This BToM model assumes that human action is a rational process and it infers the most likely intention that best explains the action trajectory so far. The output of this inference process is the posterior probability of each destination, representing the intention that the agent is likely pursuing. Figure 4 shows that, compared to the no-observer experiment (Cheng et al., 2023), participants revealed their intentions faster from 55.3% to 73.3% of their trajectories, where cluster-based permutation tests (Maris & Oostenveld, 2007) identified a significant gap (p = .007). This result indicates that humans tended to reveal their intentions faster when their actions could be observed by another person. The BToM results will be further discussed with other two experiments in General Discussion.

Experiment 2: human-AI play in parallel

In this experiment, we introduced a social context by having participants play alongside another "person" in parallel (Figure 2B). Pair of participants played the game simultaneously but in separate rooms. They were aware of each other's presence and were informed that they could observe the other player's actions on their monitors. However, the "other player" they were observing was actually an artificial intelligence (AI) agent, modeled by Markov Decision Process (MDP) (Figure 5A; this AI was a commitment-free agent who acts with a softmax policy to maximize expected utilities, Cheng et al., 2023). To emphasize the individual nature of the task, participants were explicitly informed that there was no interaction between their actions in the game; it did not matter whether they chose the same destination or not.

Method

Participants

The sample size was the same as in Experiment 1, with 50 participants (comprising 25 pairs; M_{age} = 20.98, range = 18-27) recruited from the pool at XXX University. Each participant received a payment of 10 RMB for their participation, and an additional 5 RMB for completing the task following instructions.

Design and Procedure

Pairs of participants received instructions simultaneously from an experimenter. The main task was identical to that in Experiment 1, requiring participants to reach one of the destinations as quickly as possible using the fewest steps. Participants were explicitly told that their tasks were independent: there would be no interaction or interference between their actions or rewards. They were also informed that they were free to enter the same grid and reach the same endpoint in the game. After receiving instructions, participants were placed in separate rooms to begin the main experiment. Each participant used a computer display in their respective rooms to view the game. The two computers were connected by a long cable, creating the impression that participants were seeing each other's moves on a shared screen. In reality, they were observing the actions of an MDP-AI agent. The MDP-AI's actions were synchronized with those of the human participants; as soon as a human action was received, the MDP-AI would also act, ensuring their actions appeared on the screen simultaneously.

Results

Enhanced goal perseverance

In the random disruption trials, the rationality baseline was 3.8% (95% CI: [0.019, 0.056]), which did not differ statistically from the no-observer experiment ($t(98) = 0.31$, p =

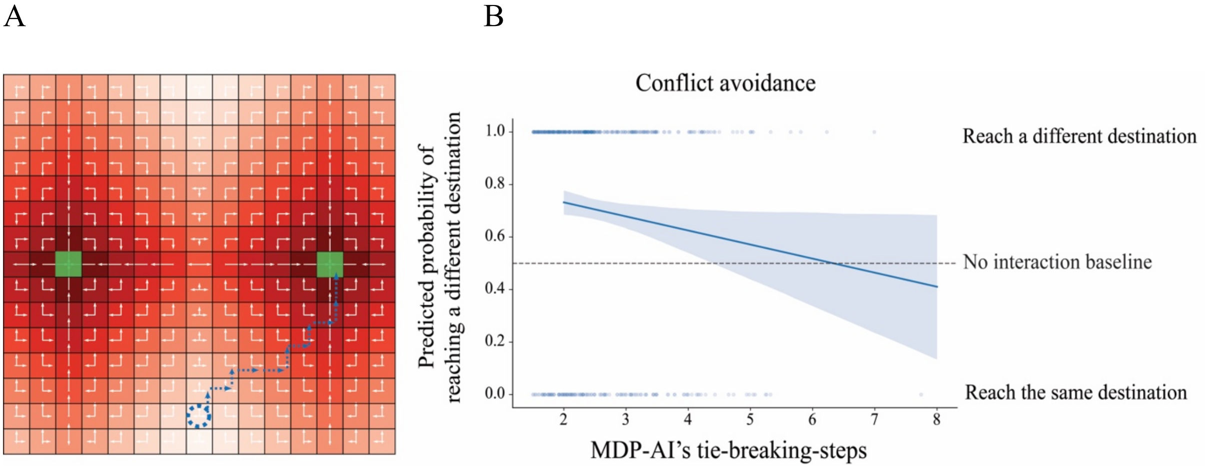
0.76, Cohen's $d = 0.06$, $BF_{10} = 0.22$). In the deliberate disruption trial, 88% of participants chose the originally-intended but later-suboptimal goal, which significantly higher than in the no-observer experiment (88% vs. 70%, Fisher's exact test, odds ratio = 3.14, one tailed $p = .0239$, Cramer's $\varphi = 0.22$; $BF_{10} = 3.29$). These results replicate the findings from Experiment 1, demonstrating that the presence of a second person can indeed lead to higher "goal perseverance".

Spontaneous intention coordination

We then explored whether the presence of a second player triggered any type of interaction. Although this was an individual task that did not require interaction, human participants might still spontaneously interpret the other agent's intentions, which could in turn influence their own intentional commitment. One possible outcome of such spontaneous interactions was a task-irrelevant "conflict avoidance," where human participants might avoid "stepping on others' shoes" by choosing different destinations from those the MDP-AI was reaching. Since the MDP-AI was a purely individual model without any consideration of the human actions, any observed interaction would be attributed unilaterally to the human participants.

Figure 5

The impact of MDP-AI's actions on human choice of destination



Note. Panel A: An illustration of the MDP-AI's softmax policy in our grid world task environment. The color of each grid represents the expected future reward starting from the current state until reaching the goal, with darker colors representing higher expected rewards. The white arrows in each grid represent the optimal policy as the distribution of actions in each state. In the states where the agent is placed in between the two destinations (green squares), there is an equal probability to move in three directions (left, right and straight). The blue dashed lines represent a trajectory with randomly sampled actions from the softmax policy. Panel B. A fitted mixed-effect logistic regression predicting "conflict avoidance" from MDP-AI's tie-breaking-steps. The error shadows reflect 95% confidence intervals.

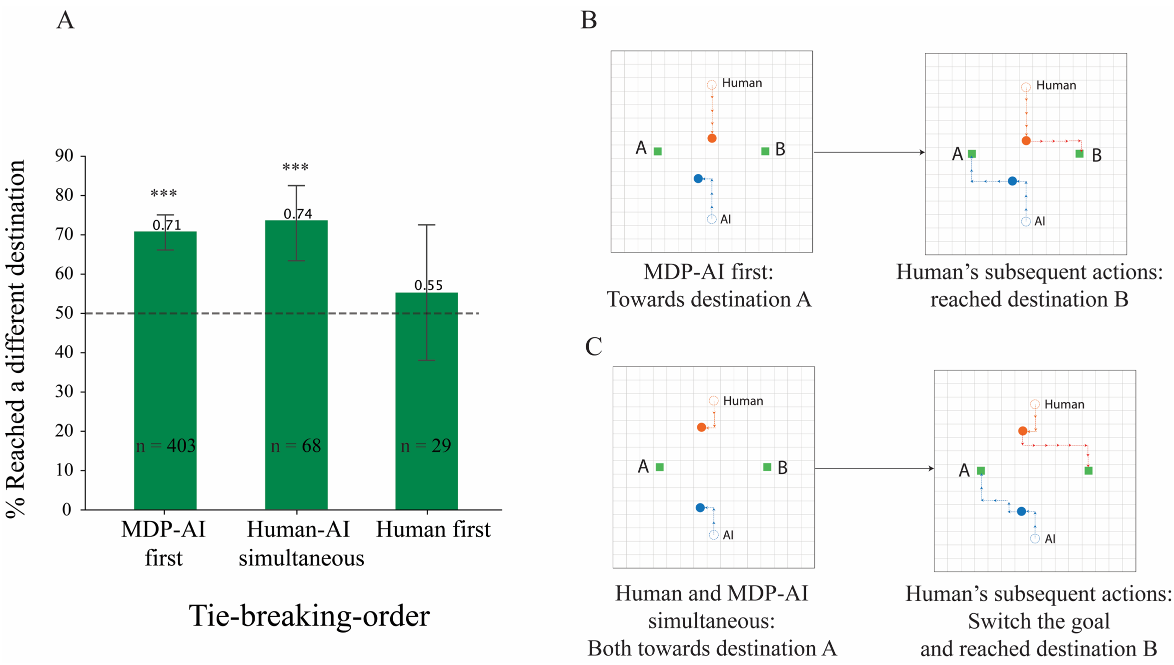
Across all trials, the "conflict avoidance" results showed that participants chose a different destination from the MDP-AI in 70.2% of the trials. A fitted mixed-effect logistic regression model, predicting the likelihood of reaching different destinations with a random by-subject intercept, revealed that this human tendency to choose different destinations occurred more frequently than the chance level of 50% ($\beta_{\text{intercept}}$ = 1.05, OR = 2.87, 95% CI [1.99, 4.15], $p < .001$). The observed "conflict avoidance" suggests that humans may spontaneously infer the destination of the MDP-AI agent, which they subsequently avoid, even if such coordination was not required for the task. As "conflict avoidance" is based on the destinations of both agents, here we further explored whether this effect is influenced by when their intended destinations were revealed respectively. We defined tie-breaking-steps as the critical step where the agent first reveals its destination—at this time, the agent is moving toward one destination while simultaneously moving away from the other. For example, in Figure 5A, the tie-breaking-steps of the MDP-AI are at steps-2.

We first explored whether "conflict avoidance" depended on MDP-AI's tie-breaking-steps. It's possible that when the MDP-AI breaks the tie quickly, humans can predict its destination earlier, therefore more likely to avoid a conflict. We conducted a mixed-effects logistic regression model to predict "conflict avoidance" from the MDP-AI's tie-breaking-steps, with random intercepts included for individual players (Figure 5B). This analysis showed a significant main effect of the MDP-AI's tie-breaking-steps (β = -0.33; OR = 0.72, 95%CI [0.57, 0.91]; *p* = .005), suggesting that the quicker the MDP revealed a destination, the more likely humans were to choose a different destination.

A plausible explanation of this temporal effect is that humans' conflict avoidance depends on which agent, MDP-AI or humans, breaks the tie first. Humans may avoid the MDP-AI's destination provided only when the MDP-AI is the first to "claim" that destination. To explore this hypothesis, we split the results into three groups based on the tie-breaking-order of the two agents: MDP-AI first, human and MDP-AI simultaneous, and human first (Figure 6A). Indeed, when the MDP-AI was the first to break the tie, which constituted the majority of the trials (80.6% of trials or 403/500), humans were more likely to choose a different destination compared to the chance level (71% or 285/403, binomial *p* < .001; Figure 6B illustrates a typical example). In the few trials where humans revealed a destination first (in 5.8% of trials, or 29/500), the percentage of reaching a different destination (55.2% or 16/29) did not significantly differ from the chance level (binomial *p* = *.71*).

Figure 6

The effect of tie-breaking-order on human's conflict avoidance



MDP-AI first:
Towards destination A

Human's subsequent actions:
reached destination B

Human and MDP-AI
simultaneous:
Both towards destination A

Human's subsequent actions:
Switch the goal
and reached destination B

Note. Panel A. Percentage of reaching a different destination based on who (human or MDP-AI) first breaks the tie. Panels B and C. Two representative trajectories of human conflict avoidance. ****p* < .001.

Interestingly, we also found that when the MDP-AI and humans revealed their destinations simultaneously (in 13.6% of trials, or 68/500), humans still tended to arrive at a different destination (73.5% or 50/68, binomial $p < .001$). This avoidance behavior, which cannot occur at the moment humans break the tie but only afterwards, indicates that humans retracted their commitments by changing their intention to a different destination, after recognizing a conflict with the MDP-AI's actions (Figure 6C). Consistently, we found that humans demonstrated high commitment to their initial choice in both the MDP-AI first and human first groups (96.77% and 93.1%, respectively; mean 96.53%), but this commitment dropped to 77.4% in the human-AI simultaneous group (96.53% vs 77.4%, Fisher's exact test, odds ratio = 0.12, one tailed $p = .0003$, Cramer's $\varphi = 0.22$; $\mathrm{BF}_{10} = 56.47$).

Overall, these results suggest that humans spontaneously apply a theory of mind to infer the MDP-AI's destination and use it to coordinate their own actions. Moreover, humans did not yield to the other agent unconditionally, but only when their intention was shown later than the MDP-AI's, suggesting a form of intention-based ownership. This interpretation will be further discussed in the General Discussion.

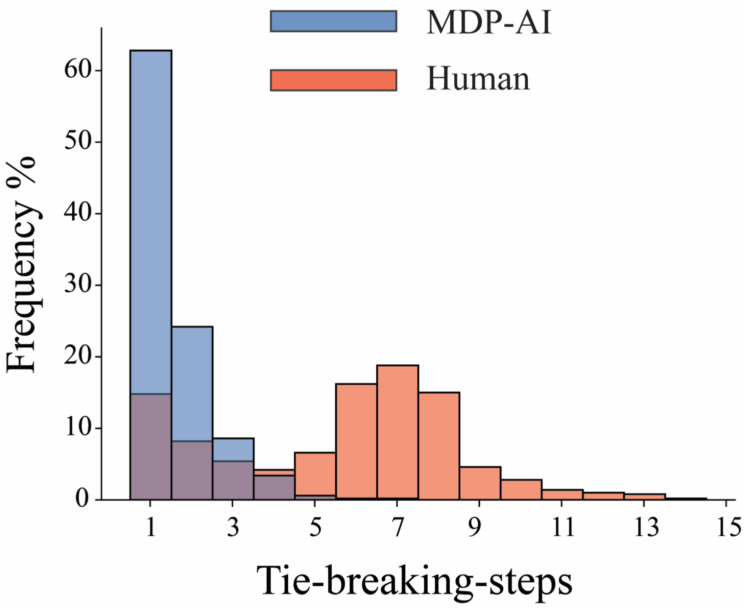
The timing of tie-breaking: Human vs. MDP-AI

The unbalanced number of trials across the three tie-breaking groups clearly shows that the MDP-AI break the tie faster than humans. This observation provides a stark contrast between the decision-making processes of humans and the MDP-AI, which lacks intention, offering valuable insights into how humans form an intention they later commit to. Here we plot the histogram of tie-breaking-steps for humans and MDP-AI to highlight these two distinctive patterns (Figure 7). Across all trials, the MDP-AI typically broke the tie within 3 steps, using an average of 1.56 steps (95%CI [1.48, 1.64]). In contrast, humans generally took many more steps, using an average of 5.56 steps (95% CI: [5.22, 5.90]). The difference between humans and MDP-AI was significant ($t(98) = 22.8$, $p < .001$, Cohen's $d = 4.56$). For the MDP-AI, the results were straightforward to explain. As shown in Figure 7, the MDP-AI followed a softmax policy. At the starting point, three actions (moving left, right, and straight) have equal expected utilities, giving each a probability of ⅓. Among these, two actions (moving left and right) will break the tie, while one (moving straight) maintains it. Therefore, within 3 steps, the probability of the MDP-AI not breaking the tie is just 1/27 (or approximately 3.7%). However, what is truly intriguing is the extended period it takes humans to broke a tie. Considering that the two destinations were always equidistant throughout the entire experiments, spending additional time to choose a destination seems unnecessary. Yet, this prolonged period suggests that humans did not break the tie randomly, but instead deliberately **maintained** the tie until a commitment to an intention is made.

Together with the findings on “goal perseverance,” the results highlight a two-stage human deliberation process: initially avoiding commitment by deliberately maintaining the tie, followed by committing to an intention once the tie is broken. We will further explore these implications in the General Discussion.

Figure 7

The timing of tie-breaking: Human vs. MDP-AI



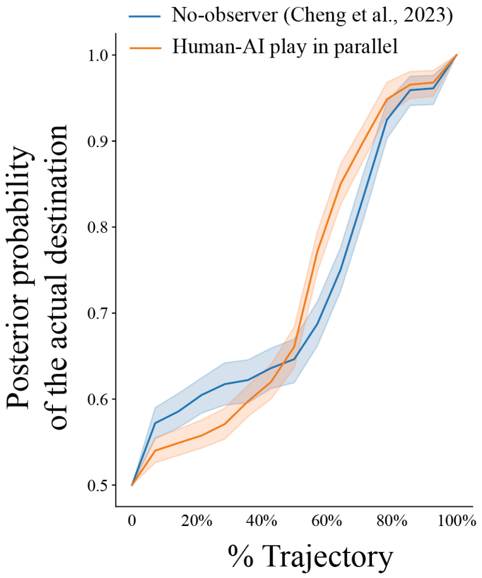
Note. Histogram of human and MDP-AI's tie-breaking-steps.

Human intention in the eyes of a Bayesian observer

The impact of the MDP-AI on human decision-making can be further revealed by applying BToM to infer the posteriors of human intentions over time, as compared to the no-observer experiment (Cheng et al., 2023). One factor contributing to the prolonged human decision process might be that humans were waiting for the MDP-AI's decision, to avoid potential conflicts. This hypothesis was supported by the BToM analysis of trajectories from all trials (Figure 8), showing that initially, humans revealed their intentions more slowly during navigation (cluster-based permutation tests identified a significant gap from 6.7% to 33.3% of the trajectories, *p* = .041). After that, humans sped up their revealing of intentions, as evidenced by a notable increase from 55.3% to 73.3% of trajectories (cluster-based permutation tests, *p* = .005). However, one must be cautious in generalizing the above findings of human tendencies, as they might be specific to interactions with the MDP-AI, which acts very differently from humans. These differences are highlighted by (1) the MDP-AI always broke a tie quickly using random sampling and (2) the self-centered nature of the MDP-AI, which ignores human actions as irrelevant to its own reward. Practically, due to this certain type of MDP-AI's actions, there were very few trials where a human player broke the tie first or simultaneously with the MDP-AI. We aimed to address these limitations in a subsequent experiment with two human players playing in parallel.

Figure 8

Human intentions in the eyes of a BToM observer



Note. The posterior of the BToM inference for the agent's actual destination as a function of the agent's steps over time. The error shadows reflect 95% confidence intervals.

Experiment 3: human-human play in parallel

This experiment mirrored the design of Experiment 2, with the key difference being that the second agent in the game was actually controlled by a real human player.

Method

Participants

The sample size and recruiting procedure for this experiment were consistent with Experiment 2. We recruited 50 participants (comprising 25 pairs; M_{age} = 20.44, range = 18-25) from the pool of XXX University. Each participant received a payment of 10 RMB for participating and an additional 5 RMB for completing the task following instructions.

Design and Procedure

The procedure for this experiment was identical to that of Experiment 2, except that after the instructions, participants played the game in a single room using one computer equipped with two monitors and two keyboards. The participants sat opposite each other on either side of a large table, each facing their respective monitor. The two monitors were wired together, displaying identical content. In this experiment, each participant's actions were updated in real time on both screens. Experiment 3 was preregistered at https://osf.io/gzvf3/?view_only=9c51337dec84440dad5d199679bc4111

Results

Enhanced goal perseverance

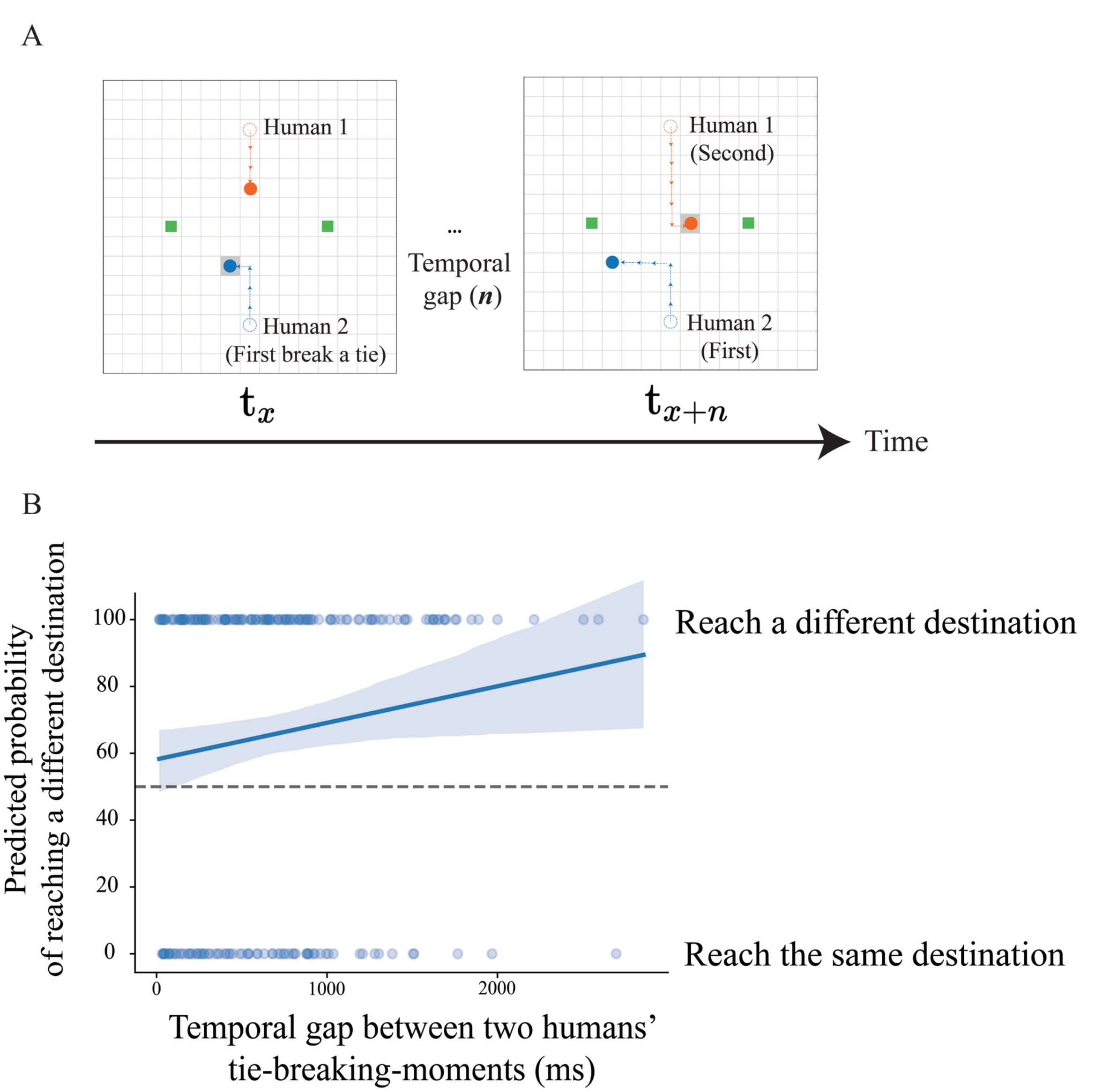
In the random disruption trials, the rationality baseline was 2% (95% CI: [0.003, 0.037]), which was not statistically different from those of the no-observer experiment (t(98) = 1.62, p = .11, Cohen's d = 0.32, BF_{10} = 0.68), Experiment 1 (t(98) = 1.54, p = 0.13, Cohen's d = 0.31, BF_{10} = 0.60), or Experiment 2 (t(98) = 1.43, p = 0.16, Cohen's d = 0.29, BF_{10} = 0.52). In the deliberate disruption trial, participants were more likely to choose the originally-intended but later-suboptimal destination compared to participants in the no-observed experiment (92% vs. 70%, Fisher’s exact test, odds ratio = 4.93, one tailed p = .0047, Cramer's φ = 0.28; BF_{10} = 14.73). This result replicates the findings of Experiments 1 and 2, indicating that the presence of a second person can indeed lead to higher "goal perseverance". No significant difference was found in "goal perseverance" between all three social contexts experiments (all two-sided ps > .5).

Spontaneous intention coordination

Across all trials, the "conflict avoidance" results showed that participants finally reached a different destination from the MDP agent in 66% of the trials. A fixed-effect logistic regression model, predicting the likelihood of reaching different destinations with a random by-subject intercept, revealed that this human tendency to reach different destinations occurred more frequently than the chance level of 50% ($\beta_{\text{intercept}} = 0.75$, OR = 2.12, 95% CI [1.4, 3.2], $p < .001$). We then measured the “temporal gap” between the two player’s tie-breaking-moments (Figure 9A), to explore how conflict avoidance was influenced by when the two humans revealed their intentions. We conducted a mixed-effect logistic regression to predict conflict avoidance from the temporal-gap, with each individual included as a random effect. This analysis revealed a significant positive coefficient for the temporal-gap ($\beta = 0.68$, OR = 1.98, 95% CI [1.10, 3.57], $p = .023$), suggesting that a larger temporal-gap leads to stronger conflict avoidance. The fitted curve in Figure 9B suggests that the longer the temporal gap, the more likely it is that the second players, who reveal their intentions later, will take into account the intentions of the first player and are more likely to choose a different destination to avoid conflict. Together, these results extended the findings of "conflict avoidance" from unilateral human-AI interactions to bilateral human-human interactions, suggesting that humans spontaneously coordinate their actions based on inferred intentions regarding the other's destination.

Figure 9

Spontaneous intention coordination



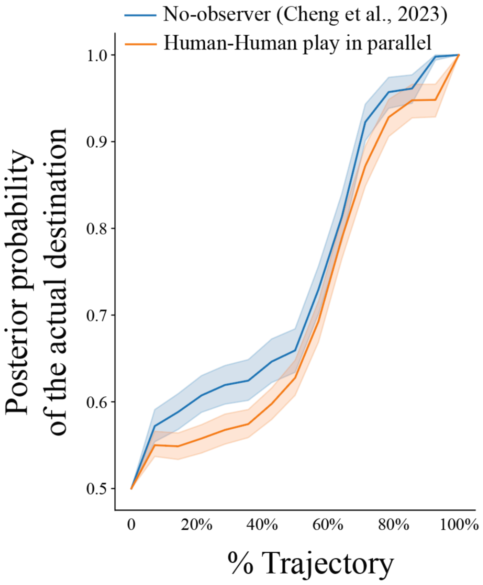
Note. Panel A. Temporal gap (n) between two players' tie-breaking-moments. Panel B. Predicting "conflict avoidance" from the temporal gap between two humans' tie-breaking-moments. The error shadows reflect 95% confidence intervals.

Human intention in the eyes of a Bayesian observer

Across all trials, the BToM analysis showed that (Figure 10), compared to the no-observer experiment, participants revealed their intentions more slowly at the beginning of their navigation. Cluster-based permutation tests found a significant difference between 13.3% to 46.7% of the trajectories (p < .001). This period of delayed intention revelation suggests that humans may exert additional effort to coordinate their intentions with each other.

Figure 10

Human intentions in the eyes of a BToM observer



Note. The posterior of the BToM inference for the agent's actual destination as a function of the agent's steps over time. The error shadows reflect 95% confidence intervals.

General Discussion

This study presents a social origin of individual commitment, demonstrating how social context can influence its strength, content, and timing. We identified two stages in human intentional actions: a commitment avoidance stage followed by a commitment stage. We found that social contexts can influence both stages, albeit in different ways.

Enhanced commitment in the presence of others

By using "goal perseverance" as a measure of commitment strength after human intention formation, we found that different manipulations of social contexts consistently lead to stronger commitment. In contrast, these social contexts generally did not impact the overall task performance, indicating that the observed enhancement in "goal perseverance" was a specific effect of social context on commitment, rather than a mere low-level arousal effect. Our work connects to previous research showing that the presence of others can promote self-presentational behaviors that have a positive social impact on others (Barclay, 2013; Engelmann & Rapp, 2018; Asaba & Gweon, 2022). Our findings further reveal that individuals are sensitive to social context even in pure individual tasks devoid of direct social impact. They spontaneously increase their commitment when acting in a public context, despite this commitment not directly contributing to task performance or social reputation. These findings are consistent with the SIE hypothesis, suggesting that due to the importance of commitment in social interaction, it has been internalized as a form of mental scheme. Thus, commitment manifests spontaneously even in individual tasks and can be automatically enhanced by the presence of others.

Spontaneous coordination via intention-based ownership

We also found that social contexts trigger spontaneous intention coordination processes, impacting both the content and timing of commitment. The phenomenon of "conflict avoidance" indicates that individuals spontaneously (1) infer the goals of others and

(2) avoid pursuing the same goals, even when their tasks are entirely independent. Furthermore, we found that "conflict avoidance" was not a universal tendency, but depended on subtle temporal dynamics of the presentation of intention between the two players. Specifically, it is the individuals who revealed a destination later that tended to avoid the earlier revealer's destination, rather than vice versa. Additionally, the tendency to avoid conflicts became stronger when the temporal gap between the two players' intention-revealing moments increased, likely due to the later revealer having more time to take the earlier revealer's intention into consideration.

Such systematic conflict avoidance, achieved without communication, highlights that common ground, a social infrastructure originally proposed in the context of cooperation involving joint commitment, is also crucial for individual commitment. Specifically, common ground refers to a shared conceptual space, mutually acknowledged and transparent (Tomasello, 2010). Our results indicate that humans spontaneously applied theory of mind to interpret each other's actions, integrating inferred intentions into their common ground, thereby facilitating conflict avoidance through the mutually recognized intention inferences. Furthermore, our findings align with the notion of respecting ownership (Friedman et al., 2008; Rossano et al., 2011), which in our case is intention-based, where the first person to present an intention is perceived as "owning" that intended destination, and therefore others should avoid that owned destination.

From avoidance to presentation of commitment: two opposing stages of intentional actions

Our findings also revealed an intriguing pattern in how humans form commitments. In philosophy and neuroscience, there have been rich discussions regarding how humans can break a tie when facing the equilibrium of two equally desirable choices (Ullmann-Margalit, 1977; Furstenberg et al., 2015). One hypothesis suggests that humans, like MDP-AI, may simply break ties through random sampling. Yet, our findings indicate that human intention formation seems more sophisticated than mere random selection, as evidenced by humans' prolonged periods of maintaining the tie. Contrary to MDP-AI, which acts without commitment throughout its trajectories, human intentional actions display two distinct deliberation phases: initially avoiding commitment to any option, followed by committing to a chosen intention. While human decision-making seems unnecessarily sophisticated compared to the MDP-AI, it is reasonable when considering the importance and consequences of making a commitment in real life. These findings suggest that one may not simply assume that humans, being goal-directed agents, will quickly manifest their intentions. When they have not decided which intention to commit to, they tend to conceal or delay revealing their intentions, an intriguing phenomenon that deserves future studies.

Different social contexts impact the intention formation process differently

While the strength of commitment was enhanced across all social contexts after an intention was formed, our BToM analysis further showed that different social contexts had different impacts on the process of intention formation. We found that the presence of a mere outside observer could speed up humans' demonstration of commitments (Experiment 1). In contrast, a second player presented in a shared visual display tended to delay such demonstrations (Experiments 2 and 3), making people more cautious in forming and revealing their intentions. This delay may arise due to a spontaneous intention coordination process being triggered, compelling the individual to invest additional effort in making a commitment. As our manipulations of social context highlight the richness behind the dynamic of intentional commitment, future studies may leverage real life social contexts as a window to further uncover the process of intention formation.

Table 1

Table of Limitations

Dimension	Assessment
	Internal validity
Is the phenomenon diagnosed with experimental methods?	Yes
Is the phenomenon diagnosed with longitudinal methods?	No
Were the manipulations validated with manipulation checks, pretest data, or outcome data?	Yes. We used outcome data to check whether participants completed the task following the instructions.
	Statistical validity
Was the statistical power at least 80%?	Yes.
Was the reliability of the dependent measure established in this publication or elsewhere in the literature?	Yes, our measures of spontaneous commitment were similar to those in the literature.
	Generalizability to different methods
Were different experimental manipulations used?	Yes. We manipulated three different social contexts to measure the impact of social presence on individual commitment.
	Generalizability to field settings
Was the phenomenon assessed in a field setting? Are the methods artificial?	The phenomenon was assessed in a video game-like task within a lab-based environment. We used a simulated game environment in which conflicting desires were represented as visual shapes with clearly defined and identical rewards. The applicability of our findings to real-world conflicting goals (e.g., working in the lab vs playing video games) warrants further investigation.
	Generalizability to times and populations
Are the results generalizable across populations (e.g., different ages, cultures, or nationalities)? Are the results generalizable to different years and historic periods?	The participants in our studies were adults in China. Although we discussed the "social inner eye" as a universal cognitive process, it is possible that different social environments and cultures (e.g., Eastern versus Western) may have different socialization pathways for shaping individual commitment, which could also lead to different developmental trajectories (e.g., Chernyak et al., 2019). Thus, it is important for future studies to explore the developmental and cultural aspects of the "social inner eye" hypothesis.

Reference

Ajzen, I., & Kruglanski, A. W. (2019). Reasoned action in the service of goal pursuit. *Psychological Review, 126*(5), 774–786. https://doi.org/10.1037/rev0000155

Ajzen, I., Czasch, C., & Flood, M. G. (2009). From intentions to behavior: Implementation intention, commitment, and conscientiousness. *Journal of Applied Social Psychology*, 39(6), 1356–1372. https://doi.org/10.1111/j.1559-1816.2009.00485.x

Arkes, H. R., & Ayton, P. (1999). The sunk cost and Concorde effects: Are humans less rational than lower animals? *Psychological Bulletin, 125*(5), 591–600. https://doi.org/10.1037/0033-2909.125.5.591

Asaba, M., & Gweon, H. (2022). Young children infer and manage what others think about them. *Proceedings of the National Academy of Sciences*, 119(32), e2105642119. https://doi.org/10.1073/pnas.2105642119

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

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist, 44*(9), 1175–1184. https://doi.org/10.1037/0003-066X.44.9.1175

Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behavior, 34*(3), 164–175. https://doi.org/10.1016/j.evolhumbehav.2013.02.002

Bratman, M. (1987). *Intention, plans, and practical reason*. Harvard University Press.

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

Chernyak, N., Kang, C., & Kushnir, T. (2019). The cultural roots of free will beliefs: How Singaporean and US Children judge and explain possibilities for action in interpersonal contexts. *Developmental Psychology*, 55(4), 866–876.

Chu, J., & Schulz, L. (2022). " Because I want to": Valuing goals for their own sake. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, No. 44).

Elkind, D. (1967). Egocentrism in adolescence. *Child Development, 38*(4), 1025–1034. https://doi.org/10.2307/1127100

Elster, J. (1987). *The multiple self*. Cambridge University Press.

Engelmann, J. M., & Rapp, D. J. (2018). The influence of reputational concerns on children's prosociality. *Current Opinion in Psychology, 20,* 92–95. https://doi.org/10.1016/j.copsyc.2017.08.024

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual review of psychology*, 50(1), 21-45. https://doi.org/10.1146/annurev.psych.50.1.21

Friedman, O., & Neary, K. R. (2008). Determining who owns what: Do children infer ownership from first possession? *Cognition, 107*(3), 829–849. https://doi.org/10.1016/j.cognition.2007.12.002

Furstenberg, A., Breska, A., Sompolinsky, H., & Deouell, L. Y. (2015). Evidence of change of intention in picking situations. *Journal of Cognitive Neuroscience*, 27(11), 2133-2146. https://doi.org/10.1162/jocn_a_00842

Goffman, E. (1959). *The presentation of self in everyday life*. Doubleday.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, 54(7), 493–503. https://doi.org/10.1037/0003-066X.54.7.493

Gollwitzer, P. M., & Brandstätter, V. (1997). Implementation intentions and effective goal pursuit. *Journal of Personality and Social Psychology*, 73(1), 186–199. https://doi.org/10.1037/0022-3514.73.1.186

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. In M. P. Zanna (Ed.), *Advances in experimental social psychology,* Vol. 38, pp. 69–119). Elsevier Academic Press. https://doi.org/10.1016/S0065-2601(06)38002-1

Gräfenhain, M., Behne, T., Carpenter, M., & Tomasello, M. (2009). Young children's understanding of joint commitments. *Developmental Psychology, 45*(5), 1430–1443. https://doi.org/10.1037/a0016122

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

Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science, 344*(6190), 1–6. https://doi.org/10.1126/science.1243091

Humphrey, N. (2002). *The inner eye*. Oxford University Press.

Kushnir, T. (2022). Imagination and social cognition in childhood. *Wiley Interdisciplinary Reviews: Cognitive Science*, 13(4), e1603.

Mack, A., & Rock, I. (1998). *Inattentional blindness*. The MIT Press

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods*, 164(1), 177-190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Mead, G.H. (1934). *Mind, Self, and Society from the Standpoint of a Social Behaviorist.* University of Chicago Press: Chicago.

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

Molinaro, G., & Collins, A. G. (2023). A goal-centric outlook on learning. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2023.08.011

Nenkov, G. Y., & Gollwitzer, P. M. (2012). Pre- versus postdecisional deliberation and goal commitment: The positive effects of defensiveness. *Journal of Experimental Social Psychology, 48*(1), 106–121. https://doi.org/10.1016/j.jesp.2011.08.002

Rochat, P. (2009). *Others in mind: Social origins of self-consciousness.* Cambridge University Press. https://doi.org/10.1017/CBO9780511812484

Rossano, F., Rakoczy, H., & Tomasello, M. (2011). Young children's understanding of violations of property rights. *Cognition*, 121(2), 219-227. https://doi.org/10.1016/j.cognition.2011.06.007

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

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences, 18*(4), 186–193. https://doi.org/10.1016/j.tics.2014.01.006

Siposova, B., Tomasello, M., & Carpenter, M. (2018). Communicative eye contact signals a commitment to cooperate for young children. *Cognition, 179,* 192–201. https://doi.org/10.1016/j.cognition.2018.06.010

Staw, B. M. (1976). Knee-deep in the Big Muddy: A study of escalating commitment to a chosen course of action. *Organizational Behavior & Human Performance, 16*(1), 27–44. https://doi.org/10.1016/0030-5073(76)90005-2

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

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

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5), 675–735. https://doi.org/10.1017/S0140525X05000129

Ullmann-Margalit, E., & Morgenbesser, S. (1977). Picking and choosing. *Social research*, 757-785.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (Cole, M., John-Steiner, V., Scribner, S. & Souberman, E., Eds.). Harvard University Press. (Original work [ca. 1930-1934])

Zhai, S., Cheng, S., Moskowitz, N., Shen, M., & Gao, T. (2023). The development of commitment: Attention for intention. *Child Development*. https://doi.org/10.1111/cdev.13955

Zhao, X., Zhao, X., Gweon, H., & Kushnir, T. (2021). Leaving a choice for others: Children's evaluations of considerate, socially-mindful actions. *Child Development*, 92(4), 1238–1253. https://doi.org/10.1111/cdev.13480