Dyadic Theory of Mind profile shapes verbal cooperation and the trajectory of non-verbal cooperation

Roksana Markiewicz [1,2], Ali Mazaheri[1,2], Katrien Segaert[1,2,3]

[1] School of Psychology, University of Birmingham, UK, [2] Centre for Human Brain Health, University of Birmingham, UK, [3] Centre for Developmental Science, University of Birmingham, UK

Corresponding author: r.markiewicz@bham.ac.uk

Data availability: Stimuli, Python scripts for the programmed experiments, and data are available here: https://osf.io/m8ut2/

Conflict of interest: The authors declare no conflicts of interest.

Acknowledgments: We would like to thank the students who helped to conduct the data collection: Sophia Taliaferro, Oscar Courtney, Shi Miao, Ffion Midgley

Abstract

Theory of Mind (ToM) is thought to support cooperation, yet it is studied as a property of individuals, even though cooperation is achieved by two people together. We examined whether cooperation depends on the ToM profile of a dyad: the pair's average ToM and the mismatch (absolute difference) between partners. Unacquainted pairs completed two cooperation tasks. In a verbal task, each partner held information needed to solve a shared problem, and partners had to talk to combine it. In a non-verbal task, partners synchronised button presses without communicating, adjusting their timing based on feedback. ToM was assessed using the Movie for the Assessment of Social Cognition (MASC; Dziobek et al., 2006); fluid intelligence was controlled. Across both paradigms, we found converging evidence that cooperative performance was shaped by the interaction between partners' ToM abilities, beyond individual ToM ability in isolation. In the verbal task (N=171 pairs), larger ToM mismatches were associated with poorer cooperative performance in pairs with relatively low average ToM, whereas higher average ToM attenuated this detrimental mismatch effect. In the non-verbal task (N=175 pairs), the same pattern emerged only over time: pairs with large ToM mismatches and relatively low average ToM improved their synchrony more slowly, so that their disadvantage was most pronounced late in the task, when cooperation depends more on anticipating the partner's adjustments. Taken together, our findings suggest that ToM operates not only as an individual socio-cognitive capacity but also as a dyadic resource that shapes how partners establish shared representations in cooperative settings.

Key words: Theory of Mind, cooperation, social cognition, social interaction, interpersonal coordination

Public significance statement

This study found that that across verbal and non-verbal tasks, successful cooperation depended on the social-cognitive profile of the pair rather than on either person in isolation. This suggests that understanding other people's thoughts and intentions operates not only as an individual ability but also as a shared resource that helps people build effective working relationships. Taken together our findings highlight the importance of considering the combination of partners' socio-cognitive abilities when understanding the mechanisms underlying successful human cooperation.

Introduction

Humans routinely coordinate and cooperate in everyday life, from subtly adjusting their walking paths to avoid collisions to engaging in complex joint activities such as performing in a musical ensemble or working on a surgical team. The effectiveness of cooperation depends on a range of factors, among which social cognition plays a central role. Social cognition encompasses processes such as Theory of Mind (ToM) or mentalizing, which is the ability to infer and understand the thoughts, intentions, and beliefs of others (Frith & Frith, 2012). In the present study, we investigated whether individual differences in ToM abilities influence cooperative performance. Given the conceptual overlap between ToM and cooperation, both of which rely on anticipating and interpreting others' mental states, we expected that indeed ToM abilities would affect cooperative behaviour. Instead of focusing on individuals in isolation, we examined how ToM abilities at the dyadic level shape cooperative performance, considering both the pair's overall level of ToM ability and the degree to which the two individuals are similar or different in this ability.

The need for this type of study arises from the interactive nature of cooperation, which inherently involves two or more individuals. As such, understanding cooperative behaviour requires approaches that capture the contributions of all participants rather than focusing on individuals in isolation. This presents a methodological challenge: experimental designs need to be sensitive to the dynamic nature of interaction while still allowing for reliable measurement of the independent variables (in this case ToM) and quantifiable cooperative outcomes. The present study seeks to balance experimental control with ecological validity. We use paradigms that capture real-time cooperative processes while enabling precise assessment of how ToM abilities shape both performance and its development over time.

Furthermore, we examined whether this effect is domain-specific, presenting differently in verbal versus non-verbal cooperative contexts, or domain-general, such that it affects both forms of cooperation in a similar way. ToM was assessed using the Movies for Assessment of Social Cognition (MASC) task (Dziobek et al., 2006), which is a well-established measure of ToM in the young adult population. Verbal cooperation was measured using a previously developed symbol matching task (Markiewicz, Rahman, et al., 2024) and non-verbal cooperation was examined using a temporal synchrony task previously used in an EEG hyperscanning context (Markiewicz, Segaert, et al., 2024).

Verbal cooperation

Verbal communication is one of the most intricate cognitive processes that humans acquire. However, successful communication is not solely dependent on language production and comprehension abilities but also on creating conceptual alignment (Stolk et al., 2016) between conversation partners. Accounts of verbal communication such as the bilateral framework proposed by Clark and Krych (2004) conceptualize speaking and listening as inherently joint processes that require continuous coordination between interlocutors. Within this framework, successful verbal communication depends on establishing mutual belief. Mutual belief is the shared understanding that each participant has sufficiently comprehended the other for the purposes of the shared goal. This process is central to verbal cooperation and forms the basis of the verbal cooperation paradigm used in the present study. The process relies on monitoring one's partner during mechanisms such as attending to speech, identifying and interpreting utterances, establishing common ground, and formulating appropriate responses. Disruptions to this partner-monitoring process have been shown to reduce communicative efficiency (Clark & Krych, 2004). This is related to theories that conceptualize verbal communication as a cooperative process towards a shared situation model (Garrod & Pickering, 2009; Pickering & Garrod, 2004, 2006).

The current study adopts a task used previously, in which cooperation is assessed within pairs of participants (Markiewicz, Rahman, et al., 2024). Each partner is given access to only part of the information necessary to complete the task. Partners are therefore required to effectively and verbally communicate with one another and integrate the information from both partners' perspectives to achieve a shared goal. Although participants are presented with different pieces of visual information, their roles are not strictly defined. This contrasts with paradigms such as maze games (Garrod & Anderson, 1987) or the Director's Task (Keysar et al., 2000), where communication is typically structured around a clear sender–receiver dynamic. Instead, in our task both individuals contribute actively without predefined roles or a fixed dialogue structure. The interaction unfolds as an open-ended conversation aimed at establishing mutual understanding, which in turn makes the task more closely resemble everyday communicative cooperation.

We extend the paradigm introduced by Markiewicz, Rahman, et al. (2024) in two key ways. First, we increase ecological validity by examining cooperation in in-person interactions rather than online exchanges conducted via Zoom. Second, we aimed to investigate cooperative learning effects and cooperative trajectories over time by maintaining a stable time limit across all trials. Cooperative performance is operationalised as trial-by-trial accuracy. These design choices enabled us to track how cooperative performance evolves across repeated interactions and to examine how these trajectories may differ as a function of pairs' ToM abilities (i.e., both, pair's average ability and between-partner mismatches).

Non-verbal cooperation

Non-verbal cooperation, often referred to as joint action, differs from verbal forms of social interaction as it does not rely on linguistic exchange. Instead, it involves the coordination of actions between two or more individuals across time and space without a verbal exchange,

for example passing someone in a narrow shopping aisle. Successful joint action depends on the formation of shared representations as well as the ability to predict both one's own actions and those of others, including making appropriate adjustments when outcomes deviate from expectations (Sebanz et al., 2006). These predictive processes can be understood in terms of the what, where, and when (Sebanz & Knoblich, 2009). The what concerns the type of action and its underlying goal, the where relates to the spatial context in which the interaction unfolds, and the when refers to the precise timing necessary for effective coordination. Together, these components enable individuals to align their behaviour and avoid interference or cooperative error.

There are several core mechanisms that underpin successful joint action (or non-verbal cooperation). First, shared attention allows individuals to focus on the same objects or events, providing a common reference point. Second, action observation facilitates the formation of shared representations of goals and supports the prediction of others' behaviour. Finally, and perhaps most critically for the non-verbal paradigm used in the present study, action adjustment enables individuals to dynamically adapt their behaviour in both time and space to accompany that of their partner. Together, these processes support the flexible and continuous coordination required for successful non-verbal cooperation.

In the current study, non-verbal cooperation was assessed using a time-estimation paradigm (adapted from a task used previously in Markiewicz, Segaert, et al. (2024). In the task, pairs of participants were instructed to synchronise button presses by establishing a shared understanding of timing intervals (short, medium, long) without verbal communication. Importantly, trial-by-trial feedback helped participants to adjust their responses across trials, allowing us to examine not only instantaneous cooperative performance but also the development of cooperative strategies over time. We thus examined whether pairs' ToM

(both, pair's average ability and between-partner mismatches) predicted cooperative performance and its trajectory across the trials.

Cooperation and Theory of Mind

Despite the theoretical importance of ToM for social interaction, empirical evidence linking ToM to cooperation remains mixed. Some studies suggest that stronger ToM abilities facilitate cooperative behaviour (Paal & Bereczkei, 2007), aligning with accounts proposing that mentalising skills are shaped through repeated social interactions (Langley et al., 2022). Further, Vogel et al. (2026) demonstrated that individuals higher in empathy (closely related to the affective dimension of ToM) are more likely to engage in prosocial behaviour and perceive helping as less costly.

However, our previous work challenges purely individual-level accounts by emphasising the inherently dyadic nature of cooperation. From this perspective, cooperative outcomes emerge from the interaction between partners' cooperative abilities rather than from individual traits alone. Accordingly ToM should be conceptualised at the dyadic level reflecting the combined and interactive contributions of both partners rather than being treated solely as an individual attribute. Consistent with this view, our previous work demonstrates that ToM abilities in both members of a dyad jointly predict cooperative failure in interactive tasks (Markiewicz, Rahman, et al., 2024). Pairs where both partners' ToM ability was high made significantly fewer cooperative errors than pairs in which both partners showed low ToM competency. These findings point to the importance of moving beyond individual-difference frameworks toward models that capture the dynamic nature of social cognition in cooperative contexts.

Aim

The present study aims to examine whether cooperative outcomes are shaped by the dyadic ToM profile, including the interaction between pair's average ToM ability levels and the

degree of ToM mismatch between partners. We use complementary verbal and non-verbal cooperation paradigms. We also aim to assess cooperative learning trajectories over time and test whether ToM-related effects generalise across modalities.

Methods

Participants

We recruited 400 healthy young adults (200 pairs) across two tasks. We aimed to recruit a sample comparable to our previous study where we also recruited 200 pairs (Markiewicz, Rahman, et al., 2024). Of these, 200 pairs completed the verbal cooperation task and 197 pairs completed the non-verbal task. Table 1 displays the core characteristics of the participants for each task.

For the verbal task, 29 (i.e., N = 58) pairs were excluded due to: limited English fluency (N = 4), misunderstanding of task instructions (N = 4), technical issues (N = 12), missing MASC data (N = 2), being friends or partners within a dyad (N = 20), low control MASC scores (>2 SD below the mean, <15; N = 5), or missing Raven's Progressive Matrices data (N = 3) along with their partners (N = 8). The final sample used for the analysis was 171 pairs (i.e., N = 342).

For the non-verbal task, 22 (i.e., N = 44) pairs were excluded due to: technical issues (N = 2), missing MASC data (N = 4), being friends or partners within a dyad (N = 20), low control MASC scores (>2 SD below the mean, <15; N = 7), or missing Raven's Progressive Matrices data (N = 2) along with their partners (N = 9). The final sample used for the analysis was 175 pairs (i.e., N = 350).

Retrospectively, this present sample provides 80% power ($\alpha = .05$) to detect an effect of the size reported in our previous study ($\eta p^2 = .055$), which requires 170 pairs and which our present sample meets (171 pairs and 175 pairs respectively). Based on this, we conclude that our new sample is large enough to detect the effect observed previously.

A sensitivity analysis (which does not rely on previously observed effect sizes) indicates that, with 171 pairs for the verbal task, our study has 80% power ($\alpha = .05$) to detect pair-level effects of $f^2 \geq 0.048$ (partial $r \geq .21$) for individual predictors, including the ToM average × ToM mismatch interaction. With 175 pairs for the non-verbal task, sensitivity for pair-level effects was comparable. The power of the 3-way interaction (ToM average × ToM mismatch × trial) was not formally established as this would require simulation-based power analyses.

Table 1. *Demographic characteristics of participants completing the verbal and non-verbal cooperation tasks.*

	Verbal task	Non-verbal task
N	342 (171 pairs)	350 (175 pairs)
Age	(20.06, SD =2.65, max =35, min = 18)	19.96 (SD = 2.54), max = 35, min = 18
Gender	252 females, 81 males, 7 non-binary/ third gender individuals whilst 2 people preferred not to state	258 females, 82 males, 8 non-binary/ third gender individuals whilst 2 people preferred not to state
English as first language	296 (221 of which were monolingual speakers)	304 (226 of which were monolingual speakers)
Control ToM	18.14 (SD = 1.29)	18.10 (SD = 1.28)
ToM	34.60 (SD = 3.72)	34.45 (SD = 3.85)

Raven's Progressive Matrices	44.86 (SD = 6.32)	44.81 (SD = 6.33)

All participants included in the subsequent analyses reported that they were not previously acquainted with their assigned partner. Participants were assigned a partner by sign-up availability. All participants had typical or corrected-to-typical vision, typical hearing, and no reported history of neurological or language impairments. All participants were fluent in English (see Table 1 for the distribution of monolingual and bilingual speakers).

Participants were students at the University of Birmingham and received either Amazon vouchers or course credit in exchange for their participation. Prior to the study, all participants received an information sheet and provided written informed consent. The study was approved by the Science, Technology, Engineering, and Mathematics Ethical Review Committee for the University of Birmingham (Ethics Approval Number: ERN_19-1661).

Materials

Verbal cooperation task

To measure verbal cooperation, we employed the experimental paradigm previously reported in Markiewicz, Rahman, et al. (2024) with some modifications. The task involved two complementary roles, which were arbitrarily assigned to two players within a pair at the beginning of the experiment and alternated after each block, ensuring that both participants completed each role for half of the session.

The task was based on the idea that each partner had access to different pieces of information, and they were required to talk to each other to combine what they knew in order to solve the task. Participants interacted freely and continuously, mirroring real-life cooperative communication.

On each trial, Player 2 was presented with six columns of symbols, each containing eight symbols (see Figure 1B and 1D). The symbols were randomly drawn from a pool of 120 unique items. One column served as the target column, while the remaining five were distractors. The identity of the target column was unknown to Player 2. Simultaneously, Player 1 was shown four symbols randomly selected from the target column on Player 2's screen (see Figure 1A and 1C). Player 1's task was to describe these four symbols in sufficient detail to enable Player 2 to identify the correct column. After identifying the target column, Player 2 was required to verbally report the order of the target symbols as they appeared from top to bottom within that column, omitting any distractor symbols. Player 1 then selected the symbols in the order recited by Player 2. Once a symbol was selected, a blue border appeared around it and selections could not be undone. Participants were informed of this at the start of the experiment as part of their instructions. A trial was considered successful if all symbols were selected in the correct order within the time limit. At the end of each trial (or when time expired), both participants received visual feedback: a green smiley face indicated successful cooperation, a red sad face indicated incorrect performance, and an orange sad face indicated that the time limit had run out.

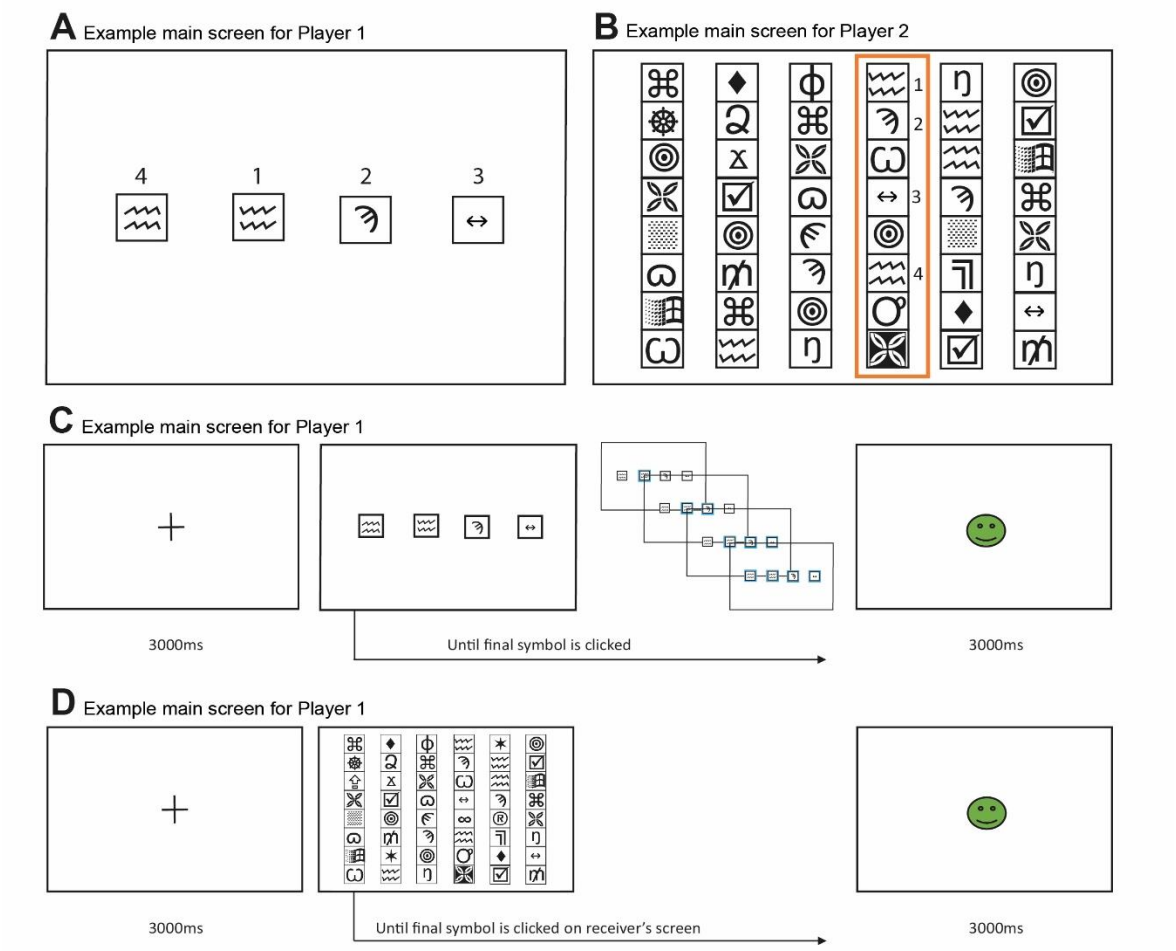

Figure 1. Example stimuli display and trial sequence for Player 1 (panels A and C) and Player 2 (panels B and D) of the verbal cooperation task. In each trial, Player 1 described the symbols shown on their screen. Based on this description, Player 2 identified the column on their screen that contained all the symbols described by Player 1. Player 2 then recited the symbols in the order they appeared in the column from top to bottom, omitting any distractor symbols. Player 1 subsequently selected the symbols in the order provided by Player 2. For clarity, the target column is highlighted in orange, and the numbers above/ adjacent to the symbols reflect the order in which Player 1 should select them. Note that the final screen of the trial sequence (panels C and D) depicts the feedback display. In this example, the trial was successful; thus, participants were shown a green smiley face. Unsuccessful trials were followed by a red sad face, whereas trials in which the time limit ran out before completion were followed by an orange sad face.

The task comprised 80 trials divided into eight blocks of 10 trials each. Roles alternated after each block. Each trial had a fixed duration of 20 seconds. A red countdown bar appeared during the final 10 seconds and progressively decreased in size each second. We made this modification relative to the previous study (Markiewicz, Rahman, et al., 2024). The previous version had a variable time limit that went down as the task progressed. However, this did not allow us to examine the learning rate/ learning trajectory. Keeping the time limit stable across the whole task allowed us to assess how cooperative behaviour develops over time and how this ties in with ToM competency.

Each trial resulted in one of three outcomes: (1) successful cooperation (correct order within the time limit), (2) failed cooperation (incorrect order), or (3) time ran out. Note that together these three outcomes accounted for 100% of the trials (trials were not limited to correct and incorrect responses alone). However, consistent with the previous study we focus on successful and failed cooperation trials only. We do not report the analyses of time-out trials as we do not consider this outcome theoretically informative. It is not possible to determine whether participants were progressing toward a correct solution when time elapsed.

Non-verbal cooperation task

To assess non-verbal cooperation, we adapted an experimental paradigm previously reported in an EEG hyperscanning study by (Markiewicz, Segaert, et al., 2024), excluding the EEG component. The task involved two participants with equivalent roles who were required to synchronize their button presses based on auditory cues.

At the beginning of each trial, both participants simultaneously received an auditory stimulus via Sennheiser HD 280 headphones. Three types of tones were used: a high-frequency tone (2500 Hz), a medium-frequency tone (1000 Hz), and a low-frequency tone (200 Hz). These tones indicated the duration participants should wait before pressing their button in an

attempt to synchronize with their partner. Specifically, participants were instructed to wait a short interval after the high-frequency tone, a medium interval after the medium-frequency tone, and a long interval after the low-frequency tone. Crucially, neither participant knew how their partner internally represented these temporal categories (short, medium, long). No verbal or non-verbal communication was permitted during the task and participants could only infer their partner's timing strategy based on trial-by-trial performance feedback.

Each trial began with the presentation of a fixation cross at the center of the screen, which remained visible throughout the trial. After 300 ms, the auditory cue was presented for 200ms. Participants were required to withhold their response for the duration corresponding to the cue and then press their button to synchronize with their partner. Once both participants responded, the fixation cross remained on the screen for an additional 1500ms, followed by feedback presentation.

Cooperation was classified as successful when the two button presses occurred within 250ms of each other. If the interval between button presses exceeded 250ms, the trial was categorized as failed cooperation. For successful trials, both participants received visual feedback in the form of a green smiling face. Failed trials were indicated by a red sad face. This visual feedback was displayed for 1000ms, followed by a fixation cross for 1000ms. Subsequently, detailed feedback was presented for 5000ms. In successful cooperation trials, participants were presented with a 'Well done!' message. In failed cooperation trials, the feedback specified which participant responded first and reported the exact temporal difference between responses.

An example trial sequence is shown in Figure 2. The experiment comprised 210 trials divided into 7 blocks of 30 trials each. Each block contained an equal number of the three auditory cues (10 high, 10 medium, and 10 low frequency tones) presented in randomized order.

Across the entire experiment, each auditory cue was presented 70 times. Participants were offered a break after each block.

The experiment was programmed in Python 3.6 using in-house scripts and functions from PsychoPy. The task scripts are publicly available at https://osf.io/m8ut2/. Both participants viewed identical instructions, stimuli, and trial sequences on separate monitors (resolution: 1920 × 1080). Responses were recorded using a Razer gaming keypad equipped with high-precision mechanical switches and internal circuitry supporting a polling rate of up to 1000 Hz, enabling millisecond-level response time registration.

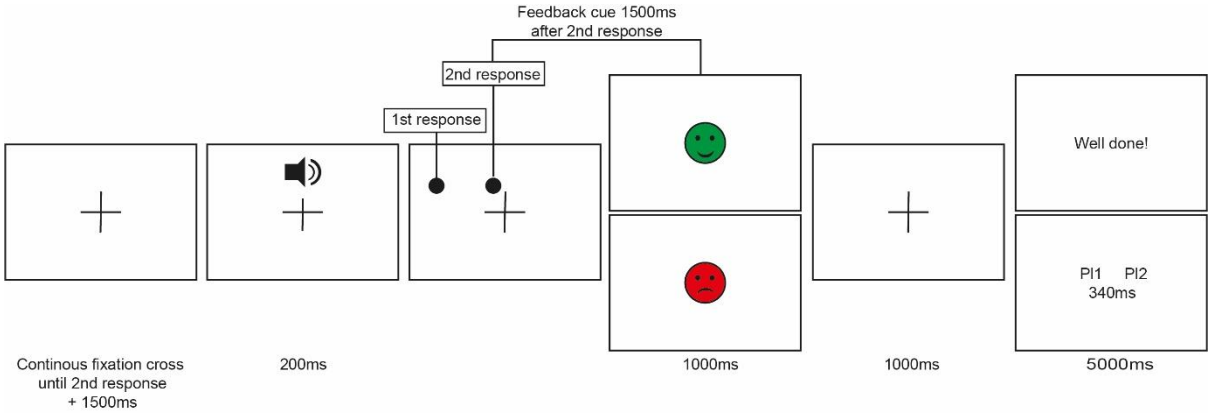

Figure 2. Example trial sequence and timing for both players of the non-verbal cooperation task. Both players were presented with the same stimuli simultaneously on two separate but identical monitors. Participants listened to the auditory cues through headphones and used a gaming pad to record their responses. The order of the button responses was completely driven by the participants thus resulting in trial-to-trial variations of who pressed first/second. Each trial was followed by feedback stimuli. Successful trials were followed by a green smiley face and “Well done!”, and unsuccessful trials were followed by a red sad face and detailed feedback text outlining which player pressed the button first/ second and the elapsed time between the button presses. Based on this feedback, players were able to learn and adapt their responses in future trials. Participants were aware of their player number from the start of the experiment so they could understand the feedback text.

Theory of Mind task

Theory of Mind (ToM) was assessed using the Movie for the Assessment of Social Cognition (MASC; Dziobek et al., 2006) task, administered online via Qualtrics (Provo, Utah, United States). In the MASC task, participants watched a series of 46 short video clips that together formed a coherent story involving four characters. After each clip, participants answered a multiple-choice question probing their understanding of the characters' mental states, including thoughts, feelings, and intentions.

To ensure that participants sustained attention to the narrative, 21 control questions were interspersed randomly throughout the task. These questions assessed factual aspects of the clips (e.g., "Which chips does Betty have to play?") rather than requiring mental state inferences. Performance on these items served as an index of attentional engagement with the task. Participants who scored more than 2 SD below the group mean (i.e., 15) on the control questions were excluded from analysis.

Each mental state question offered multiple response options and was scored as either an appropriate or an insufficient ToM response. Insufficient responses can be further categorized into hypermentalizing (over-interpretation of mental states), undermentalizing (reduced mental state attribution), and no mentalizing (absence of mental state reasoning) (e.g., Hatkevich et al., 2019). However, for the purposes of the present study, we only analysed appropriate ToM responses. The total ToM score reflected the number of appropriate responses, with higher scores indicating more accurate mental state reasoning. Possible scores ranged from 0 to 46.

Fluid intelligence

When examining individual differences in Theory of Mind and cooperation, it is important to control for general fluid intelligence to ensure that any observed effects are not simply driven by broader cognitive ability. Accordingly, we assessed non-verbal fluid intelligence via Raven's Progressive Matrices (Raven, 1958, 2000) administered online via Qualtrics. The task comprised 60 visual pattern problems. Each item had one missing segment and participants were required to select the option that correctly completed the pattern from a set of multiple-choice responses. Items were presented in order of increasing difficulty. Participants were given a total of 20 minutes to complete the task. Performance was indexed by the total number of correctly solved items within the time limit.

Procedure

All participants read the information sheet and signed a paper version of the consent form. The study consisted of two parts. In part one, participants completed two cooperation tasks: one verbal and one non-verbal. Task order was counterbalanced such that half of the participants completed the verbal cooperation task first and the non-verbal task second, while the other half completed them in the reverse order.

Part two involved individual testing and included the MASC and the Raven's Progressive Matrices tasks.

During part one, participants were seated in the same room with an occluder placed between them. This was to prevent participants from using visual cues such as facial expressions (during either task). Each participant was seated approximately 60 cm from a monitor on which the tasks were presented. Participants were encouraged to take short breaks between blocks and between the two cooperation tasks.

Upon completion of part one, participants were moved to a different room to complete the MASC and Raven's Progressive Matrices individually. This replicated the procedure of the original study. A compulsory minimum three-minute break was imposed between the MASC and Raven's Progressive Matrices tasks. Part one lasted approximately 90 minutes and part two lasted approximately 45 minutes.

Data preparation

We removed trials that had deviations from the experimental procedure or poor data quality. Specifically, in the verbal cooperation task there were two reasons for trial exclusions. Firstly, both participants used separate computer mice connected to a single laptop, with one participant designated as Player 1 on each trial to control the cursor. Participants were instructed that if they were not Player 1, they should not touch their mouse as doing so would interfere with the other participant's cursor movements and potentially affect performance. Nevertheless, some participants reflexively handled the mouse. In cases where this disrupted task performance, the trial was documented by the experimenter and excluded from analysis.

Additionally, some trials were affected by participants misunderstanding the task instructions. This most often occurred during the first few trials or when player roles changed. In these cases, the experimenter repeated the instructions during the trial. These trials were documented and excluded from analysis. On average, 2.04 trials were removed per pair (SD = 5.34), leaving an average of 77.96 valid trials per pair (SD = 5.34). Trial removal was performed manually based on the experimenter's session notes.

In the non-verbal cooperation task, we removed trials based on response times according to three criteria: (1) either player failed to press their button (mean removed trials = 1.21, SD = 6.20), (2) either player responded in less than 200 ms from trial onset (mean removed trials = 0.47, SD = 3.47), and (3) either player's response time exceeded 3 standard deviations

above/below the mean for that pair and condition (mean removed trials: short = 2.44, SD = 1.12; medium = 1.59, SD = 1.10; long = 1.94, SD = 1.29; overall per pair = 5.95, SD = 2.35). Across all criteria, an average of 7.41 trials per pair (SD = 7.37) were excluded, leaving an average of 202.59 valid trials per pair (SD = 7.37). Trial removal was performed using a custom Python script.

Results

To examine how each pair's Theory of Mind (ToM) ability and the mismatch in ToM ability between partners influenced cooperation (both verbal and non-verbal), ToM was operationalised at the pair level in two ways. We calculated: (1) the pair's average ToM ability (operationalised as the mean ToM score of both partners within a pair with higher values indicating better overall ToM ability of the pair; *ToM average*), and (2) the mismatch in ToM ability between partners (operationalised as the absolute difference in ToM scores between partners within a pair with higher values indicating greater mismatch within the pair; *ToM mismatch*).

Further, to control for general cognitive ability in our models, we applied the same approach to the Raven's Progressive Matrices scores. Specifically, we calculated: (1) the pair's average general cognitive ability (operationalised as the mean Raven's Progressive Matrices score of both partners within a pair with higher values indicating better overall general cognitive ability of the pair; *Raven average*) and (2) the mismatch in general cognitive ability between partners (operationalised as the absolute difference in Raven's Progressive Matrices scores between partners within a pair with higher values indicating greater mismatch within the pair; *Raven mismatch*). All of the predictor variables mentioned above (ToM average, ToM

mismatch, Raven average, and Raven mismatch) were mean-centred for ease of interpretation.

Verbal cooperation task

Verbal cooperative performance was indexed in the following ways: (1) successful cooperation (i.e., the proportion of trials with correct responses), (2) failed cooperation (i.e., the proportion of trials with incorrect responses), and (3) time ran out (i.e., the proportion of trials where participants did not respond within the time limit for the trial). Separate models were fitted for successful and failed cooperation. We do not report on time ran out trials (see Methods section for justification).

We examined whether pairs' average ToM ability and mismatches in ToM ability between partners not only influenced overall cooperative performance, but also whether these effects developed over the course of the experiment. To investigate this, we conducted linear mixed-effects model analyses at the pair level using the *lmer* function in R.

Three mixed-effects models were fitted for successful and failed cooperation separately. In Model 1, the fixed effects included ToM average, ToM mismatch, their interaction and Block. Model 2 extended Model 1 by additionally including Raven average, Raven mismatch, and their interaction as fixed effects. Model 3 further extended Model 2 by including Block, ToM average and ToM mismatch interactions. All models included pair as a random intercept. Note that block was treated as a continuous predictor to capture the trajectory of cooperative learning over time rather than estimating separate effects for each block.

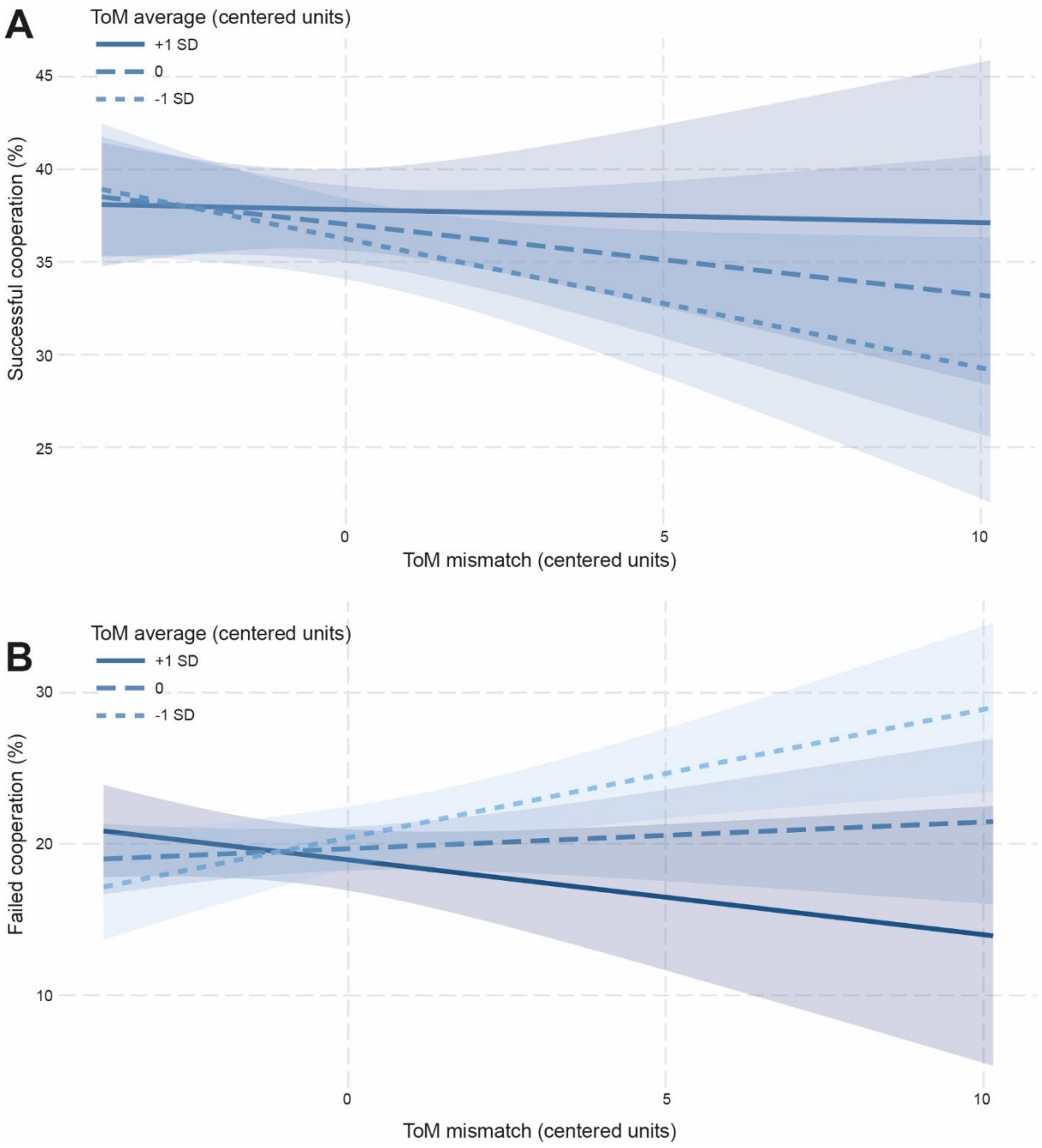

Figure 3. Cooperative performance as a function of Theory of Mind (ToM) average and ToM mismatch between partners for successful (**A**) and failed (**B**) cooperation. Lines represent predicted values from the linear mixed-effects model, with shaded areas indicating 95% confidence intervals. Separate lines represent low (-1SD; dotted line), average (mean; dashed line), and high (+1SD; solid line) levels of pair ToM average. The results indicate that the association between ToM mismatch and cooperative performance depends on the pair's average ToM. Greater ToM mismatches are associated

with significantly lower successful and higher failed cooperation primarily among pairs with lower average ToM.

Successful cooperation

Model comparisons predicting successful cooperation over blocks indicated that Model 2 provided the best fit to the data. A likelihood ratio test showed that adding the Raven predictors (Raven average, Raven mismatch, and their interaction) significantly improved model fit relative to Model 1, $\chi^2(3) = 14.253$, $p = .003$. In contrast, extending the model to include higher-order interactions between block and the ToM predictors (Model 3) did not significantly improve model fit compared to Model 2, $\chi^2(3) = 0.830$, $p = .842$. Supplementary Table 1 displays full results from all three models.

The final best-fitting model (Model 2) was re-estimated using restricted maximum likelihood to obtain unbiased estimates of the random effect variance. The model included block, ToM average and ToM mismatch, Raven average and Raven mismatch, and their respective interactions as fixed predictors (Successful cooperation ~ Block + ToM average * ToM mismatch + Raven average * Raven mismatch), with pair as a random intercept. Table 2 displays the full results of the model.

The model showed that successful cooperation increased across blocks, indicating that partners were able to develop effective cooperation strategies over time. In addition, pairs with higher average ToM and higher average Raven scores showed greater cooperative success, highlighting the importance of both socio-cognitive and domain-general cognitive abilities at the pair level. Critically, the significant interaction between ToM average and ToM mismatch shows that the impact of mismatches in ToM on successful cooperation depends on the overall level of ToM within the pair, with greater ToM average buffering

against the detrimental effects of ToM mismatch on cooperation. Random intercept variance for pair was 0.013 (SD = 0.112), with a residual variance of 0.025 (SD = 0.160), indicating modest variability between pairs.

Table 2. *Fixed Effects Estimates for successful cooperation (i.e., proportion of trials with correct responses) from the final re-estimated Mixed Effects Model. The model includes block, ToM average and ToM mismatch, Raven average and Raven mismatch, and their respective interactions as fixed predictors with a random intercept for pairs. P-values <.05 are highlighted in bold.*

Predictor	**Estimate**	**SE**	**t-value**	**p-value**
N = 1339, pairs = 171				
Intercept	0.3684	0.0105	35.023	**< .001**
Block	0.0471	0.0019	24.402	**<.001**
ToM average	0.0079	0.0038	2.10	**.037**
ToM mismatch	−0.0038	0.0036	−1.07	.287
ToM average × ToM mismatch	0.0031	0.0012	2.64	**.009**
Raven average	0.0083	0.0023	3.66	**< .001**
Raven mismatch	0.0020	0.0022	0.90	.368
Raven average × Raven mismatch	−0.0003	0.0004	−0.75	.454

Since the best-fitting model did not include the higher-order interaction between ToM average, ToM mismatch, and block, this suggests that Theory of Mind (ToM) does not differentially influence the development of successful cooperation over time, but rather the overall level of successful cooperation. A model including this three-way interaction (Model 3 reported in the Supplementary Material) confirmed that this effect was not significant. A

similar but mirrored pattern was observed for failed cooperation (see Supplementary Material), indicating ToM-related effects modulate cooperative learning trajectories in this task

As shown in Figure 3A, slopes analyses showed that among pairs with relatively low average ToM (-1SD), greater ToM mismatches between partners were associated with significantly lower successful cooperation (b = -0.01, SE = 0.00, t = -2.94, p = <.001). In contrast, this association was not significant for pairs with mean (b = -0.00, SE = 0.00, t = -1.08, p = .28) or high average ToM ability (b = 0.00, SE = 0.01, t = 0.90, p = .37). This was further supported by the results from Johnson-Neyman analysis revealing that ToM mismatches significantly predicted successful cooperation only for pairs with ToM average below -0.97 (centered units).

Together, these findings suggest that ToM mismatches between partners impair successful cooperation primarily when the pair's average ToM is relatively low. As the pair's average ToM ability increases, this detrimental effect diminishes and is no longer statistically detectable. Specifically, mismatches in ToM are most detrimental when average ToM is low, whereas higher average ToM appears to buffer against the negative consequences of mismatched ToM abilities between partners.

Failed cooperation

Similarly to successful cooperation, we examined whether pairs' ToM ability and mismatches in ToM ability between partners not only influenced overall cooperative performance, but also whether these effects developed over the course of the experiment in failed cooperation.

Model comparisons predicting failed cooperation (i.e., error rates) across blocks indicated that Model 1 provided the best fit to the data. A likelihood ratio test showed that adding the

Raven predictors (Raven average, Raven mismatch, and their interaction) in Model 2 provided a trend-level improvement relative to Model 1, $\chi^2(3) = 6.89$, $p = .075$. Extending the model to include higher-order interactions between block and the ToM predictors (Model 3) did not improve model fit compared to Model 2, $\chi^2(3) = 1.940$, $p = .586$. Additionally, Model 3 did not fit the data better than Model 1, $\chi^2(6) = 8.829$, $p = .183$. Supplementary Table 2 displays full results from all three models.

The final best-fitting model (Model 1) was re-estimated using restricted maximum likelihood to obtain unbiased estimates of the random effect variance. The model included block, ToM average and ToM mismatch, and interaction between ToM predictors as fixed predictors (Failed cooperation ~ Block + ToM average * ToM mismatch), with pair as a random intercept. Table 3 displays the full results of the model.

Table 3. *Fixed Effects Estimates for cooperative failure (i.e., proportion of trials with erroneous responses) from the final re-estimated Mixed Effects Model. The model includes block, ToM average and ToM mismatch, and their interaction as fixed predictors with a random intercept for pairs. P-values <.05 are highlighted in bold.*

Effect	**Estimate**	**SE**	**t-value**	**p-value**
N = 1339, pairs = 171				
(Intercept)	0.1967	0.0073	26.86	**< .001**
Block	0.0051	0.0016	3.11	**.002**
ToM average	−0.0026	0.0026	−1.00	.320
ToM mismatch	0.0017	0.0026	0.68	.500
ToM average × ToM mismatch	−0.0024	0.0008	−2.92	**.004**

The model showed that failed cooperation (i.e., error rates) increased across blocks. Neither ToM average nor ToM mismatch significantly predicted failed cooperation as main effects. However, the interaction between ToM average and ToM mismatch significantly predicted failed cooperation, indicating that the combined influence of pair average ToM and ToM mismatch affected failed cooperation (i.e., errors). Specifically, larger mismatches in ToM between partners are associated with greater failed cooperation when the pair's average ToM is low. In contrast, when average ToM is high, this negative effect is attenuated implying that higher average ToM acts as a buffer, allowing pairs to maintain low levels of failed cooperation despite mismatches between partners. Random intercept variance for pair was 0.006 (SD = 0.077), with a residual variance of 0.018 (SD = 0.135).

As shown in Figure 3B, simple slopes analyses showed that among pairs with relatively low average ToM (-1SD), greater ToM mismatches between partners were associated with significantly higher rates of failed cooperation (i.e., errors) (b = 0.01, SE = 0.00, t = 2.87, p = <.001). In contrast, this association was not significant for pairs with mean (b = -0.00, SE = 0.00, t = 0.68, p = .50) or high average ToM ability (b = 0.00, SE = 0.00, t = -1.27, p = .21). This was further supported by the results from Johnson-Neyman analysis revealing that ToM mismatches significantly predicted failed cooperation only for pairs with ToM average below -1.32 and above 5.39 (centered units), although the upper region represented only one pair.

These findings mirror those from successful cooperation, suggesting that higher ToM mismatches between partners affect cooperative behaviour overall when the pair's average ToM ability is relatively low. However, as the pair's average ToM ability increases, this association diminishes, suggesting that higher average ToM ability may buffer against the consequences of mismatched ToM abilities between partners. This is reflected in both successful and failed cooperation.

Non-verbal cooperation task

In the non-verbal cooperation task, we aimed to analyse the temporal component of the synchronized button presses, rather than focusing solely on the binary accuracy of cooperation. Although trials were classified online as successful or unsuccessful based on a 250ms threshold, examining temporal synchrony provides a more continuous and fine-grained measure of cooperative performance. We did not focus on raw reaction time (RT) mismatches between button presses, as these do not account for systematic variations in RTs across the intermixed timing conditions (short, medium, and long). Instead, we calculated an inter-personal timelag for each trial using the formula: $|RT_1 – RT_2| / (RT_1 + RT_2)$. This measure normalises the mismatch between partners' responses relative to their overall response speed, allowing comparisons across conditions. Accordingly, inter-personal timelag was used as the primary continuous measure of cooperation, rather than raw RT mismatches. In addition, and for completeness, we also examined overall cooperative accuracy and report on this in the Supplementary Material.

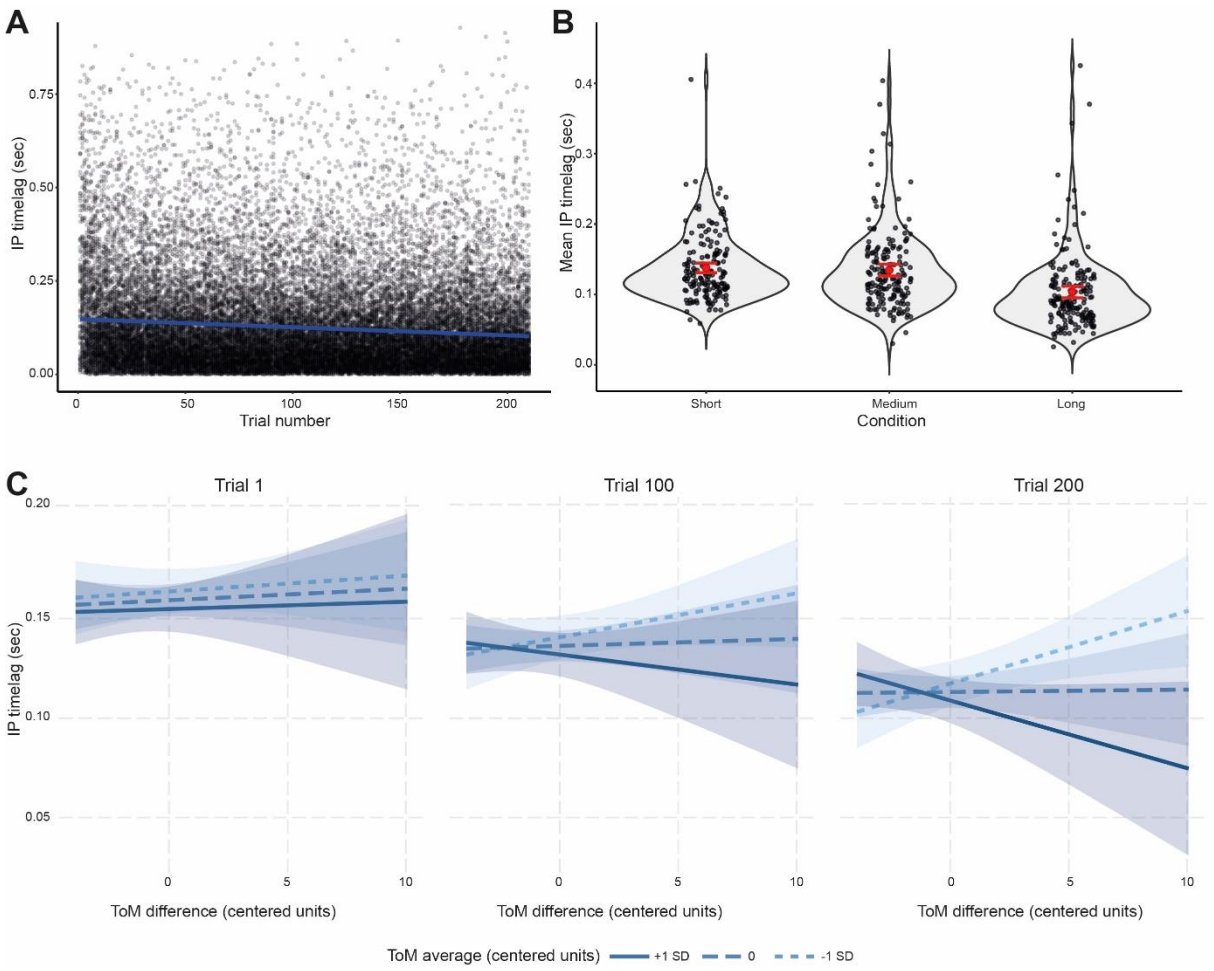

Figure 4. Inter-personal timelag results across the non-verbal cooperation task. **(A)** Scatterplot of inter-personal timelag across trials with a linear regression line representing the overall trend. Points represent individual observations and the fitted line indicates a decrease in inter-personal timelag over the course of the experiment, consistent with increased non-verbal cooperation over the course of the experiment. **(B)** Violin plots illustrating the distribution of pair-level mean inter-personal timelag across the three conditions (short for high tone, medium for medium tone, and long for low tone). Grey dots represent pair-level mean inter-personal timelag values, red points indicate condition means and error bars represent 95% confidence intervals. **(C)** Model-estimated inter-personal timelag as a function of partner Theory of Mind (ToM) mismatch and moderated by ToM average across the task trajectory (200 trials). Panels represent early (trial 1), middle (trial 100), and late (trial 200) stages of the non-verbal cooperation task (see Suppl. Figure 1 for a more detailed depiction of task-related changes across experimental trajectory, with model estimates presented for every 10th trial). Lines

represent predicted relationships between ToM mismatches and inter-personal timelag for pairs with low (−1 SD), average (mean), and high (+1 SD) ToM average ability, with shaded regions representing 95% confidence intervals. Positive slopes indicate that greater discrepancies in partners' ToM abilities are associated with increased inter-personal timelag (reduced non-verbal cooperation), whereas flatter slopes indicate a weaker association between ToM mismatch and cooperation. The influence of partner ToM mismatches changes over the course of interaction, becoming most pronounced among pairs with lower average ToM ability during later stages of the task.

Inter-personal timelag across trials

We examined whether pairs' average ToM ability and mismatches between partners not only influenced overall inter-personal timelag, but also whether these effects developed over the course of the experiment.

To investigate this, we conducted linear mixed-effects model analyses at the pair level using the *lmer* function in R. Three mixed-effects models were fitted for inter-personal timelag. In Model 1, the fixed effects included ToM average, ToM mismatch, their interaction, condition and trial. Model 2 extended Model 1 by additionally including Raven average, Raven mismatch, and their interaction as fixed effects. Model 3 further extended Model 2 by including trial, ToM average and ToM mismatch interactions. All models included pair as a random intercept. Note that trial was used as a continuous fixed effect, as we were interested in the trajectory of cooperative learning over time rather than merely accounting for the extra variability (if used as random effect).

We also included condition as a fixed factor in the models because our analyses revealed significant differences in overall accuracy (see Supplementary Table 3) and inter-personal timelag (see Supplementary Table 4) between conditions. Including condition as a fixed

effect allowed us to control for these differences across the experimental manipulations, ensuring that any effects of the ToM predictors were not confounded by condition. However, our primary interest was in the relationships between ToM measures and the inter-personal timelag over time, rather than in how these effects might differ across conditions. Therefore, we only modelled main effects of condition and did not include interactions between condition and the ToM predictors. This approach controls for condition-related variance without overcomplicating the model with interactions that were not theoretically related to our hypotheses.

A likelihood ratio test showed no significant improvement from Model 1 to Model 2 by adding the Raven predictors ($\chi^2(3) = 4.53$, $p = .210$). In contrast, Model 3 provided a significantly better fit than both Model 1 ($\chi^2(6) = 33.47$, $p < .001$) and Model 2 ($\chi^2(3) = 28.94$, $p < .001$). Therefore, Model 3 was the best-fitting model, indicating that including the three-way interaction between ToM average, ToM mismatch, and trial number substantially improved prediction of inter-personal timelag. Supplementary Table 5 displays full results from all three models.

The final, best-fitting model (Model 3) was re-estimated using restricted maximum likelihood (REML) to obtain unbiased estimates of the random effect variance. The model included trial number, condition, ToM mean, ToM mismatch, Raven mean, Raven mismatch, and their relevant interactions as fixed predictors: Inter-personal timelag ~ condition + ToM mean × ToM mismatch × trial number + Raven mean × Raven mismatch. Table 4 displays the full results of the model. The model showed that the inter-personal timelag decreased over time (i.e., the cooperative synchrony increased; see Figure 4A), with shorter inter-personal timelag in the long and medium conditions compared to the short condition (see Figure 4B). There was also a small effect of the average Raven score, such that a higher combined fluid intelligence score of a pair was associated with shorter inter-personal timelag (i.e., better

cooperative synchrony). In contrast, ToM measures did not show any clear main effects. Importantly, there was a three-way interaction between ToM average, ToM mismatch, and trial number (see Figure 4C for visualisation of this three-way interaction) indicating that the relationship between the combined ToM score of a pair and inter-personal timelag changed over time depending on ToM mismatches within pairs. Other interaction effects were not significant.

Table 4. *Fixed Effects Estimates for inter-personal timelag (calculated on a trial-by-trial basis using the following formula: $|RT_1 - RT_2| / (RT_1 + RT_2)$) from the final re-estimated Mixed Effects Model. The model includes condition, ToM mean, ToM mismatch, Raven mean, Raven mismatch, and their relevant interactions as fixed predictors with a random intercept for pairs. The reference condition was short. Trials were treated as a continuous predictor. P-values <.05 are highlighted in bold.*

Predictor	**Estimate**	**SE**	**t-value**	**p-value**
N = 35453, pairs = 175				
Intercept	0.1593	0.0041	38.797	**< .001**
ToM average	-0.0015	0.0014	-1.088	.278
ToM mismatch	0.0006	0.0013	0.436	.663
Trial number	-0.0002	0.0000	-21.194	**< .001**
Raven average	-0.0016	0.0008	-1.991	**.048**
Raven mismatch	-0.0004	0.0008	-0.478	.633
Condition (long)	-0.0344	0.0015	-22.814	**< .001**
Condition (medium)	-0.0034	0.0015	-2.269	**.023**

Predictor	Estimate	SE	t-value	p-value
ToM average × ToM mismatch	-0.0001	0.0004	-0.169	.866
ToM average × Trial number	0.00000	0.0000	0.069	.945
ToM mismatch × Trial number	-0.00000	0.0000	-0.616	.538
Raven average × Raven mismatch	-0.00007	0.0001	-0.563	.574
ToM average × ToM mismatch × Trial number	-0.00000	0.0000	-5.210	**< .001**

To examine the nature of the three-way interaction between ToM average x ToM mismatch x trial number, follow-up simple slopes analyses were conducted at early (trial 1), middle (trial 100), and late (trial 200) stages of the task (see Figure 4C and Table 5 for comprehensive results). The analyses indicated that the effect of ToM mismatch emerged only during the late stage of the task (trial 200), where greater ToM mismatch predicted increased inter-personal timelag among pairs with low average ToM ability (−1 SD; p = .01). No significant effects were observed during the early or middle stages of the task or among pairs with average or high ToM average ability.

Table 5. *Simple slopes were estimated from the three-way interaction between ToM average, ToM mismatch and trial number. The association between ToM mismatch and inter-personal timelag was examined at early, middle and late stages of the task at low (-1SD), mean, and high (+1SD) levels of ToM average. Positive coefficients indicate that greater mismatches in ToM between partners were associated with increased inter-personal timelag (reflecting poorer non-verbal cooperation).*

Task stage (trial)	ToM average	Estimate (b)	SE	t-value	p-value
Early (trial 1)	Low (−1 SD)	0.000	0.000	0.53	.60
	Mean	0.000	0.000	0.43	.66
	High (+1 SD)	0.000	0.000	0.18	.85
Middle (trial 100)	Low (−1 SD)	0.000	0.000	1.55	.12
	Mean	0.000	0.000	0.27	.78
	High (+1 SD)	−0.000	0.000	−0.79	.43
Late (trial 200)	Low (−1 SD)	0.00	0.00	2.46	**.01**
	Mean	0.001	0.00	0.09	.93
	High (+1 SD)	−0.00	0.00	−1.72	.09

Discussion

The present study examined whether cooperation is shaped by the dyadic ToM profile of interacting partners, specifically their combined (average) ToM and the degree to which the partners differ (ToM mismatch). We assessed cooperation using two complementary paradigms capturing verbal and non-verbal forms of cooperation. The verbal cooperation task required partners to communicate and integrate information to jointly solve a shared problem, whereas the non-verbal cooperation task required partners to synchronise their button presses based on auditory cues without any verbal communication, relying on trial-by-trial feedback to learn and adapt to each other's timing. Across both the verbal and non-verbal tasks, suggesting a domain-general effect, we found that cooperative behaviour is driven less by individuals in isolation and more by how partners' ToM abilities interact. In particular, large mismatches in ToM between partners within a pair tend to hinder cooperation, especially when the combined ToM ability of the pair is low. This pattern emerges across both verbal and non-verbal domains, with these dyadic effects extending beyond overall performance to the trajectory of cooperative learning in the later stages of the non-verbal task.

Together, our findings extend the view of ToM as an individual capacity that predicts social behaviour: ToM also functions as a dyadic resource, whose consequences for cooperation depend on the abilities of both partners. Successful cooperation requires individuals to make assumptions about their partner's mental states, intentions, and future actions. To do so, individuals must construct and continuously update mental models of their interaction partner and in turn use these predictions to guide their own behaviour (Manera et al., 2013; Ramnani & Miall, 2004; Springer et al., 2012; Vesper et al., 2013). Our findings extend this perspective by demonstrating that cooperation depends not only on an individual's capacity to infer others' thoughts and intentions, but also on the combined socio-cognitive profile of the dyad. Successful cooperation emerges from how well partners' cognitive abilities complement one another rather than from the abilities of either individual alone.

In the following sections, we discuss how dyadic ToM profiles influence verbal and non-verbal cooperation, respectively. Although these forms of interaction differ in their communicative demands, both rely on the same fundamental processes, namely the ability to infer a partner's internal states, predict their future behaviour, and update these prediction models to guide their own actions (Pesquita et al., 2018; Sebanz et al., 2003, 2006; Sebanz & Knoblich, 2009; Vesper et al., 2010).

Verbal cooperation

For pairs with greater ToM mismatches, relatively low average ToM was associated with poorer cooperative performance (reduced cooperative accuracy and increased cooperative errors). However, this ToM mismatch association with cooperative performance was not present for pairs with average or high overall ToM ability. In other words, pairs with strong ToM abilities were able to better compensate for mismatches between them and thus buffer against the detrimental effects of ToM partner mismatches.

The specific demands of the verbal cooperation task provide a potential explanation for the effects of the dyadic ToM profiles on cooperative performance. This task required the partners to merge complementary information that was asymmetrical in nature. Successful cooperation depended on partners developing a sufficiently shared representation of the target. This placed demands on individuals to continually represent what their partner knows, what information they have access to and what information they still need. This in turn would allow them to anticipate what their partner is likely to understand from the information they provided. These processes are essential in establishing and maintaining common ground, whereby interlocutors develop adequately shared understanding to coordinate their subsequent behaviour (Clark & Brennan, 1991). Higher-ToM partners may therefore be better able to adapt their descriptions and interpretations to their partner's perspective as well as recognise when their partner has misunderstood or requires clarification. This may facilitate faster development of conceptual alignment and reduce the likelihood of partners proceeding with divergent representations of the target (Pickering & Garrod, 2004).

In contrast, pairs characterised by both greater ToM mismatches and lower average ToM may be less efficient at establishing common ground and in turn conceptual alignment within the time constraints of the task. A mismatch in ToM ability within a pair may also create a mismatch in how partners represent and respond to each other's understanding of the situation. One partner may consider information that the other partner does not yet have or may assume a level of understanding that their partner has not reached. This can lead to greater misalignment. Once misalignment happens, interactive repair is necessary (Pickering & Garrod, 2004). Pairs with dyadic ToM profiles characterised by high ToM mismatches and low average ToM ability may commit a greater number of misalignment errors and recovery from these may be more costly and less efficient as opposed to those with high average ToM ability (Markiewicz, Rahman, et al., 2024).

.

The nature of the current dyadic ToM results is consistent with our previous study (Markiewicz, Rahman, et al., 2024) which also showed that the profile of ToM abilities within a pair shapes cooperative behaviour (specifically failed cooperation). Previously, we found that pairs in which both partners had high ToM abilities made significantly fewer cooperative errors than pairs in which both partners had low ToM abilities, while mixed pairs did not differ reliably from either group. In the current study, by modelling ToM as continuous dyadic characteristics rather than imposing arbitrary high vs. low categories, we demonstrate that both the average ToM ability of a pair and the degree of mismatch between partners jointly shape cooperative behaviour. Specifically, ToM mismatches were most detrimental to cooperation when pair-level ToM was low, suggesting that it is not simply the presence of a mismatch between partners that matters but also how this mismatch interacts with the overall socio-cognitive resources available within the dyad. These findings extend our previous work by showing that cooperation is shaped by the balance and combination of partners' ToM abilities. Unlike our previous study, this effect emerged across both successful (accuracy) and failed (error rates) cooperation, potentially reflecting the greater sensitivity of in-person interactions to partner-level dynamics compared with online testing.

However, this dyadic ToM profile does not appear to shape the trajectory of verbal cooperative learning itself. One possible explanation is limited statistical sensitivity. The experiment only included eight blocks, and thus the design may not be fine-grained enough to detect subtle differences in learning trajectories. Alternatively, ToM in the verbal task may exert a more stable and cumulative influence on overall performance rather than shaping how quickly pairs improve. Additionally, analysing performance at the block level may be too coarse to capture moment-to-moment adaptations, meaning that any ToM-related mismatches

in learning dynamics could be obscured despite the overall improvement observed across blocks.

Non-verbal cooperation

The findings from the non-verbal cooperation paradigm support and extend the conclusion from the verbal cooperation paradigm. Here, we also observe that the interaction between the pair's average ToM ability and the mismatch in ToM between partners was associated with cooperative performance. However, the results from the non-verbal task provide an additional insight by capturing the trajectory of cooperative performance over time. Specifically, the effect of ToM mismatches on cooperation (indexed by the inter-personal timelag between synchronised button presses) depended on the pair's average ToM ability with this interaction becoming most pronounced in the late stages of the task. Although pairs generally improved their non-verbal cooperation over time, pairs characterised by both relatively low average ToM and large ToM mismatches showed a reduced rate of improvement in the later trials. Thus, dyadic ToM profiles become increasingly crucial as cooperation progresses from responding to general task demands towards developing partner-specific expectations and adapting behaviour to the particular behavioural tendencies of the other individual.

One potential explanation relates to the demands of cooperation in the absence of explicit communication. In pure coordination games, individuals rely on salient solutions or intuitive alignment when communication is not available (Schelling, 1960). This may have supported the initial non-verbal cooperation in the present task, as participants could rely on the most salient interpretation of the short, medium, and long timing intervals (Perez-Zapata et al., 2021). However, while salient solutions or intuitive alignment can facilitate coordination when clear conventions exist, salient solutions alone are unlikely to be sufficient as cooperation develops. The later stages of our non-verbal task require partners to increasingly

learn how the other person responds and continuously update their predictions of their partner's behaviour.

The successful cooperation process in our task is not simply a matter of adjusting one's own response based on provided feedback but rather about adapting while simultaneously considering how the partner will also respond to that feedback. For example, if one partner responds 500 ms faster than the other on a given trial then the optimal adjustment on the next trial (of the same condition) is not simply to slow down by 500 ms in order to synchronise. Instead, each partner must anticipate that the other will also adapt their strategy based on the previous feedback. If both partners independently make the full corrective adjustment, their responses may overshoot in the opposite direction rather than converge. Effective non-verbal cooperation therefore requires a process of mutual adjustment where both partners converge towards a shared conceptual alignment strategy to be developed (Markiewicz, Segaert, et al., 2024; Stolk et al., 2016). This does not only require sensitivity to one's own performance feedback but also the ability to represent others' actions and integrate these representations with one's own action planning (Sebanz et al., 2006; Sebanz & Knoblich, 2009).

Similarly, predictive accounts of joint action propose that individuals continuously generate and update predictions about their partner's behaviour in order to coordinate their own actions effectively (Kourtis et al., 2019; Pesquita et al., 2018). From this perspective, the later emergence of the dyadic ToM profile interaction may reflect a shift from general task-based learning towards partner-specific adaptation. Early in the task, participants may primarily need to learn the task-specific demands and the required responses, whereas later, successful cooperation may depend increasingly on learning the particular behavioural tendencies of the partner and anticipating how they will respond to feedback.

This may explain why dyads with greater mismatches in ToM ability and lower average ToM showed a reduced (or even impaired) improvement in non-verbal cooperation during the later stages of the task. These pairs may have placed greater emphasis on individual ('self') feedback than on integrating self-information with partner-related predictions, which in turn limited their ability to integrate information about their own performance with predictions about their partner's behaviour. This self-focused updating may have hindered the development of shared representations required for effective non-verbal cooperation (Kourtis et al., 2019; Sebanz et al., 2003). Rather than simply learning that their own responses were too early or too late, successful cooperation required individuals to infer how their partner might interpret the same feedback and how this would influence their subsequent response. A greater mismatch in ToM ability may therefore create asymmetry in partners' capacity to generate and update such predictions. When the average dyadic ToM is relatively low, neither partner may have sufficient socio-cognitive resources to compensate effectively for this asymmetry. This in turn would make it more difficult for the dyad to converge on a stable and mutually effective cooperation strategy.

On the other hand, when average ToM ability was high, the negative impact of ToM mismatches between partners appeared to be attenuated. These dyads may be better able to use the provided feedback to modify their responses not only in accordance with individual feedback but also with their predictions of how their partner would adapt. Higher average ToM ability of a pair may therefore accommodate mismatches in partners' socio-cognitive abilities by supporting more effective partner modelling and prediction updating. Furthermore, these dyads may also be able to engage in coordination smoothing behaviours (Vesper et al., 2011), whereby individuals adapt their own behaviour in ways that make it easier for their partner to predict it and in turn facilitate cooperation. Therefore, higher

average ToM ability of a pair may buffer against the effects of ToM discrepancies by enabling partners to establish more predictable patterns of behaviour.

Future research

A future research avenue is to investigate how the neural mechanisms of cooperation relate to the socio-cognitive profile of a dyad. Our previous work (Markiewicz, Segaert, et al., 2024) demonstrated that brain-to-brain dynamics between interacting partners can predict subsequent joint action outcomes. However, an important next step would be to determine whether these neural signatures are influenced by, for example, the ToM profile of a pair. Although our current results demonstrate that dyadic ToM profile is associated with cooperation, they cannot determine the underlying mechanisms (i.e., development of prediction models, prediction updating, neural synchrony or coordination smoothing). Integrating behavioural and neural approaches would provide a more complete understanding of how interacting partners achieve successful cooperation.

Conclusion

Taken together, the current findings provide complimentary evidence that successful cooperation is shaped not merely by the socio-cognitive abilities of individuals but by how these abilities combine within a dyad. This relationship emerged in different aspects of cooperation across the two tasks. In the verbal cooperation task, ToM mismatches of a pair were detrimental to overall cooperative performance primarily among pairs with relatively low average ToM ability, whereas higher dyadic average ToM buffered against this effect. In the non-verbal task, dyadic interaction emerged across the trajectory of cooperation: the detrimental effects of low-mean, high mismatch between partners became most pronounced during later stages of the task. These findings imply that studies relating social cognition to

cooperation should characterise both partners, since the effect of one partner's ToM depends on the other's. Cooperation, in short, depends on the dyad, not only on the individual.

References

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In *Perspectives on Socially Shared Cognition* (pp. 127–149). Washington, DC: APA Books.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*, 62–81. https://doi.org/10.1016/j.jml.2003.08.004

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J. K., Wolf, O. T., & Convit, A. (2006). Introducing MASC: A movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders*, *36*(5), 623–636. https://doi.org/10.1007/s10803-006-0107-0

Frith, C. D., & Frith, U. (2012). Mechanisms of Social Cognition. *Annual Review of Psychology*, *63*(1), 287–313. https://doi.org/10.1146/annurev-psych-120710-100449

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*(1987), 181–218.

Garrod, S., & Pickering, M. J. (2009). Joint Action, Interactive Alignment, and Dialog. *Topics in Cognitive Science*, *1*(2), 292–304. https://doi.org/10.1111/j.1756-8765.2009.01020.x

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38. https://doi.org/10.1111/1467-9280.00211

Kourtis, D., Woźniak, M., Sebanz, N., & Knoblich, G. (2019). Evidence for we-representations during joint action planning. *Neuropsychologia*, *131*(April), 73–83. https://doi.org/10.1016/j.neuropsychologia.2019.05.029

Langley, C., Cirstea, B. I., Cuzzolin, F., & Sahakian, B. J. (2022). Theory of Mind and Preference Learning at the Interface of Cognitive Science, Neuroscience, and AI: A Review. *Frontiers in Artificial Intelligence*, *5*(April), 1–17. https://doi.org/10.3389/frai.2022.778852

Manera, V., Schouten, B., Verfaillie, K., & Becchio, C. (2013). Time Will Show: Real Time Predictions during Interpersonal Action Perception. *PLoS ONE*, *8*(1). https://doi.org/10.1371/journal.pone.0054949

Markiewicz, R., Rahman, F., Apperly, I., Mazaheri, A., & Segaert, K. (2024). It Is Not All About You : Communicative Cooperation Is Determined by Your Partner ' s Theory of Mind Abilities as Well as Your Own. *Journal of Experimental Psychology : Learning , Memory , and Cognition*. https://doi.org/https://doi.org/10.1037/xlm0001268

Markiewicz, R., Segaert, K., & Mazaheri, A. (2024). Brain-to-brain coupling forecasts future joint action outcomes. *IScience*, *27*(9), 110802. https://doi.org/10.1016/j.isci.2024.110802

Paal, T., & Bereczkei, T. (2007). Adult theory of mind, cooperation, Machiavellianism: The effect of mindreading on social relations. *Personality and Individual Differences*, *43*(3), 541–551. https://doi.org/10.1016/j.paid.2006.12.021

Perez-Zapata, D. I., McKenzie-Smart, X., Charest, I., & Apperly, I. (2021). *Meeting and Missing Minds: Children and Adults Use Alignment of Intuitions to Solve Pure Coordination Games*. *July 3*. https://doi.org/10.31234/osf.io/96hbp

Pesquita, A., Whitwell, R. L., & Enns, J. T. (2018). Predictive joint-action model: A hierarchical predictive approach to human cooperation. *Psychonomic Bulletin and Review*, *25*(5), 1751–1769. https://doi.org/10.3758/s13423-017-1393-6

Pickering, M. J., & Garrod, S. (2004). The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, *27*(2), 212–225. https://doi.org/10.1017/s0140525x04450055

Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, *4*(2–3), 203–228. https://doi.org/10.1007/s11168-006-9004-0

Ramnani, N., & Miall, R. C. (2004). A system in the human brain for predicting the actions of others. *Nature Neuroscience*, *7*(1), 85–90. https://doi.org/10.1038/nn1168

Raven, J. C. (1958). *Standard Progressive Matrices. Sets A, B, C, D, and E.* London: H.K. Lewis & Co. Ltd.

Raven, J. C. (2000). The Raven's Progressive Matrices: Change and Stability over Culture and Time. *Cognitive Psychology*, *41*(1), 1–48. https://doi.org/10.1006/cogp.1999.0735

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

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70–76. https://doi.org/10.1016/j.tics.2005.12.009

Sebanz, N., & Knoblich, G. (2009). Prediction in Joint Action: What, When, and Where. *Topics in Cognitive Science*, *1*(2), 353–367. https://doi.org/10.1111/j.1756-8765.2009.01024.x

Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: Just like one's own? *Cognition*, *88*(3), 11–21. https://doi.org/10.1016/S0010-0277(03)00043-X

Springer, A., de Hamilton, A. F. C., & Cross, E. S. (2012). Simulating and predicting others' actions. *Psychological Research*, *76*(4), 383–387. https://doi.org/10.1007/s00426-012-0443-y

Stolk, A., Verhagen, L., & Toni, I. (2016). Conceptual Alignment: How Brains Achieve Mutual Understanding. *Trends in Cognitive Sciences*, *20*(3), 180–191. https://doi.org/10.1016/j.tics.2015.11.007

Vesper, C., Butterfill, S., Knoblich, G., & Sebanz, N. (2010). A minimal architecture for joint action. *Neural Networks*, *23*, 998–1003. https://doi.org/10.1016/j.neunet.2010.06.002

Vesper, C., Van Der Wel, R. P. R. D., Knoblich, G., & Sebanz, N. (2011). Making oneself predictable: Reduced temporal variability facilitates joint action coordination. *Experimental Brain Research*, *211*(3–4), 517–530. https://doi.org/10.1007/s00221-011-2706-z

Vesper, C., van der Wel, R. P. R. D., Knoblich, G., & Sebanz, N. (2013). Are you ready to jump? Predictive mechanisms in interpersonal coordination. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 48–61. https://doi.org/10.1037/a0028066

Vogel, T. A., Priestley, L., Cutler, J., Hogg, T., Khalighinejad, N., Garrett, N., Apps, M. A. J., Rushworth, M. F. S., & Lockwood, P. L. (2026). Humans are more prosocial in poor foraging environments. *Nature Communications* , *17*(1). https://doi.org/10.1038/s41467-025-66880-9