AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2023, Vol. 152, No. 10, 2735–2746
https://doi.org/10.1037/xge0001419

Social Metacognition Drives Willingness to Commit

Georgia E. Kapetaniou[1, 2], Ophelia Deroy[2, 3, 4, 5], and Alexander Soutschek[1, 2]

[1] Department of Psychology, Ludwig Maximilian University Munich
[2] Department of Biology, Graduate School for Systemic Neurosciences, Ludwig Maximilian University Munich
[3] Faculty of Philosophy, Ludwig Maximilian University Munich
[4] Munich Centre for Neuroscience, Ludwig Maximilian University Munich
[5] Institute of Philosophy, School of Advanced Study, University of London

Showing or telling others that we are committed to cooperate with them can boost social cooperation. But what makes us willing to signal our cooperativeness, when it is costly to do so? In two experiments, we tested the hypothesis that agents engage in social commitments if their subjective confidence in predicting the interaction partner's behavior is low. In Experiment 1 (preregistered), 48 participants played a prisoner's dilemma game where they could signal their intentions to their co-player by enduring a monetary cost. As hypothesized, low confidence in one's prediction of the co-player's intentions was associated with a higher willingness to engage in costly commitment. In Experiment 2 (31 participants), we replicate these findings and moreover provide causal evidence that experimentally lowering the predictability of others' actions (and thereby confidence in these predictions) motivates commitment decisions. Finally, across both experiments, we show that participants possess and demonstrate metacognitive access to the accuracy of their mentalizing processes. Taken together, our findings shed light on the importance of confidence representations and metacognitive processes in social interactions.

Public Significance Statement

Everyday situations, from navigating crowded sidewalks to negotiating complex trade deals, require us to make predictions about the behavior of others. Will they be cooperative or self-interested? And how confident are we in our ability to accurately predict their actions? These questions lie at the heart of our study, which uses a combination of behavioral economics and subjective confidence measures to explore the role of metacognition in social decision-making. Our findings highlight the critical importance of not only being able to predict others' behavior but also understanding the accuracy of our predictions. By including measures of metacognition in studies of social cognition, we can gain a deeper understanding of how individuals make decisions and develop more effective strategies for navigating the social world.

Keywords: social cognition, commitment, mentalizing, confidence, metacognition

Supplemental materials: https://doi.org/10.1037/xge0001419.supp

Why do people throw surprise parties for their friends and partners? Besides the entertainment, the personal sacrifice of time or money invested in displays like parties can serve as signals that one is committed to the relationship (Yamaguchi et al., 2015). Similarly, entrepreneurs can signal to a venture capitalist that they are committed to the company's development by devoting time and exerting effort to secure seed financing (Elitzur & Gavious, 2003). Personal and business relationships are only two examples of social cooperation where individuals can benefit from considering collective rather than selfish interests. Costly commitments where an

This article was published Online First May 1, 2023.

Georgia E. Kapetaniou https://orcid.org/0000-0002-0020-1158

We kindly thank the Munich Experimental Laboratory for Economic and Social Sciences (MELESSA) for support with recruitment and data collection.

Alexander Soutschek received an Emmy Noether fellowship (SO 1636/2-1) from the German Research Foundation. Ophelia Deroy is funded by the NOMIS foundation (grant DISE). All authors declare to have no conflicts of interest.

The data and code that support the findings of this study are available on Open Science Framework (https://osf.io/eq3f2/?view_only=7ff216408a9c4e24a114f9c53c0ac268).

The first experiment was preregistered at https://osf.io/y89n3.

Georgia Eleni Kapetaniou served as lead for formal analysis, investigation, software, visualization, and contributed equally to project administration. Alexander Soutschek served as lead for funding acquisition, project administration, resources, and supervision. Georgia Eleni Kapetaniou, Ophelia Deroy, and Alexander Soutschek contributed equally to conceptualization, methodology, writing–original draft, and writing–review and editing. Georgia Eleni Kapetaniou and Alexander Soutschek contributed equally to data curation.

Correspondence concerning this article should be addressed to Georgia E. Kapetaniou, Department of Psychology, Ludwig Maximilian University Munich, Leopoldstr. 13, 80802 Munich, Germany. Email: Georgia.Kapetaniou@psy.lmu.de

agent unilaterally pays a cost to signal commitment to a particular action (e.g., cooperation) reliably improve cooperation, as shown both by formal mathematical models (Arieli et al., 2017; Han et al., 2015; Renou, 2009) and empirical research (Locey & Rachlin, 2012). Commitments can be considered as a form of communication between interaction partners. Through commitments, agents reveal their goals and intentions to interaction partners before those make their decisions (Balliet, 2010; Smith & Bliege Bird, 2005). Yet even the cost–benefit calculations that could be performed before deciding whether to commit cannot eliminate uncertainty. How does this uncertainty affect the motivation to pay the costs of unilateral commitment in strategic interactions?

Costly commitments can signal an agent's intentions to the other party (Buskens & Royakkers, 2002; Harrington & Zhao, 2012), but might not always benefit the agent, as they might be exploited by a free-riding interaction partner. To decide whether to engage in costly commitment, agents, therefore, need to reliably assess the other's intentions and predict their behavior (Brosig, 2002; Fischbacher et al., 2001; Gächter, 2006; Keser & Van Winden, 2000), an ability often referred to as mentalizing. While mentalizing matters for strategic interactions, how it matters for motivating commitment deserves to be studied. Recent computational approaches clarify that mentalizing can provide an advantage in social interactions: strategies that consider beliefs about the other's intentions prevail over even the most successful strategies that don't (Han et al., 2011, 2015; Nakamura & Ohtsuki, 2016). Furthermore, models incorporating inference on others' goals can successfully explain behavioral data in different social dilemma paradigms (Diaconescu et al., 2014; Yoshida et al., 2008).

The mental states of others, however, are not easy to decipher, and inferring what someone intends is often an uncertain business (Wu et al., 2020). As agents should cooperate only when they are sufficiently confident that their partner will cooperate rather than free-ride (Frith, 2012; Wu et al., 2020), their capacity to monitor the accuracy of their mentalizing process should affect social decisions. In other words, social metacognition matters as much as mind-reading if one needs to decide to cooperate.

While costly signaling mainly aims to reduce the receiver's uncertainty about the signaler's intentions, we hypothesize that, from the signaler's perspective, uncertainty about what they infer about the receiver's intentions is an important driving force for signaling decisions. If someone predicts that the other will cooperate but with low confidence, they would benefit from signaling—even at a cost—that they are willing to cooperate in order to motivate the other to reciprocate cooperation. If someone predicts that the other will cooperate (or defect) with high confidence, the costs of signaling may outweigh its potential benefits. As signaling one's own willingness to cooperate can lower the receiver's fear of being exploited by a free-rider, signaling should be employed predominantly when a decision-maker is unsure about the other's intentions. Indeed, computational models suggest that mutual cooperation (and thus the interaction partners' payoff) can be enhanced if agents opt for costly commitment particularly when confidence in the prediction's accuracy is low (Han et al., 2015). This argument would be consistent with recent evidence that metacognitive access to one's economic preferences moderates the willingness to precommit to long-term goals in individual decision-making (Soutschek et al., 2021; Soutschek & Tobler, 2020). Still, while metacognition is known to shape decision-making across different domains of cognition (De Martino et al., 2013; Deroy et al., 2016; Fleming et al., 2012; Soutschek & Tobler, 2020), and recent neuroimaging studies document a substantial overlap between the underlying neural mechanisms of metacognition and mentalizing (Vaccaro & Fleming, 2018; Valk et al., 2016), there is surprisingly little evidence about people's metacognitive access to the accuracy of their mentalizing processes during social interactions, and even less about how this social metacognition affects strategic social interactions.

This is the gap that the current study addresses. The role of confidence in mentalizing processes (i.e., the metacognitive access to the accuracy of intention recognition processes) in costly signaling cooperation is investigated in two independent experiments. In Experiment 1, we provide evidence that decision-makers possess metacognitive access to the accuracy of their mentalizing processes and that this metacognitive knowledge guides their commitment choices. Employing a novel paradigm where decision-makers in the prisoner's dilemma game can pay a small fee to commit to cooperation or a tit-for-tat strategy, we hypothesized that decision-makers choose costly commitments particularly when they can predict the interaction partner's actions only with low confidence. In Experiment 2, we then provide evidence for the causal impact of confidence on costly commitment by assessing the impact of experimentally lowering confidence on costly signaling. Together, our findings provide consistent evidence for the role of metacognitive access to mentalizing processes in social cooperation.

Experiment 1

Experiment 1 tested the following hypotheses: First, we tested whether the option to commit to cooperation or a tit-for-tat strategy by paying a monetary cost increases cooperation in the iterated prisoner's dilemma game, as suggested by previous findings (Locey & Rachlin, 2012). Second, by trial-by-trial measuring predictions of the other's actions and the subjective confidence in these predictions, we tested whether low confidence in predicting other's intentions leads to more commitment choices. Third, we tested whether decision-makers possess metacognitive access to the accuracy of their mentalizing processes.

Transparency and Openness

Below we report and justify our sample sizes and data exclusions, as well as all manipulations and all measures for both experiments included in this article. Data analyses were performed with R, Version 4.0.3 (R Core Team, 2020). All data and analysis codes are available at https://osf.io/eq3f2/?view_only=7ff216408a9c4e24a114f9c53c0ac268 (Kapetaniou et al., 2022). The study was approved by the local ethics committee, and the design and analysis (Experiment 1) was preregistered on the Open Science Framework (https://osf.io/y89n3).

Materials and Method

Participants

Forty-eight volunteers (24 female; 24 male; $M_{age} = 23.2$ years, $SD = 3.09$) were recruited through the participant pool of the Melessa lab at the Ludwig Maximilian University Munich, Germany. The sample size was based on an a priori power analysis (power = 80%, $\alpha = 5\%$) assuming a medium effect size (Cohen's

$d = 0.5$), which indicated that 34 participants should be sufficient to detect a significant effect. We increased this sample to 48 participants, as data collection took place in groups of eight participants in order to further increase statistical power. Exclusion criteria were previous participations in prisoner's dilemma game experiments or currently studying or having majored in economics, to avoid that participants were familiar with the prisoner's dilemma game (Frank et al., 1993). All participants gave informed written consent and received compensation for their participation, which consisted of a base fee of €12 and an additional bonus depending on their choices. The study was approved by the local ethics committee and was preregistered on the Open Science Framework (https://osf.io/y89n3).

Stimuli and Task Design

All tasks were programmed using z-Tree, Version 4.1.6 (Fischbacher, 2007).

Prisoner's Dilemma. We used a modified version of the finitely repeated prisoner's dilemma game with two anonymously matched Players A and B who in each trial, simultaneously and without knowing the other's choice, decided to cooperate or defect. If both players cooperated, they both obtained a payoff of four coins; if both players defected, they obtained only two coins. If one player cooperated and the other defected, the free-rider received seven coins and the cooperator one coin (Figure 1). We selected this particular payoff structure (which violates the second inequality commonly used in the prisoner's dilemma game) to increase the dilemma between cooperation and defection, as it makes it more profitable for Player B to defect in the short run, whereas reaching a cooperative equilibrium becomes more uncertain, thus making Player B's behavior less predictable for Player A (Krach et al., 2009).

Participants performed this task under three conditions (within-subject): In the *no commitment* condition, they played the prisoner's dilemma game as described (Figure 1A). Participants had to decide between cooperation and defection within 6 s via mouse click on the corresponding buttons presented on the top and bottom of the screen (counterbalanced across trials). At the end of each trial, they received feedback (1 s) about their and the other's payoff (e.g., "Your Profit: 4, Other's Profit: 4").

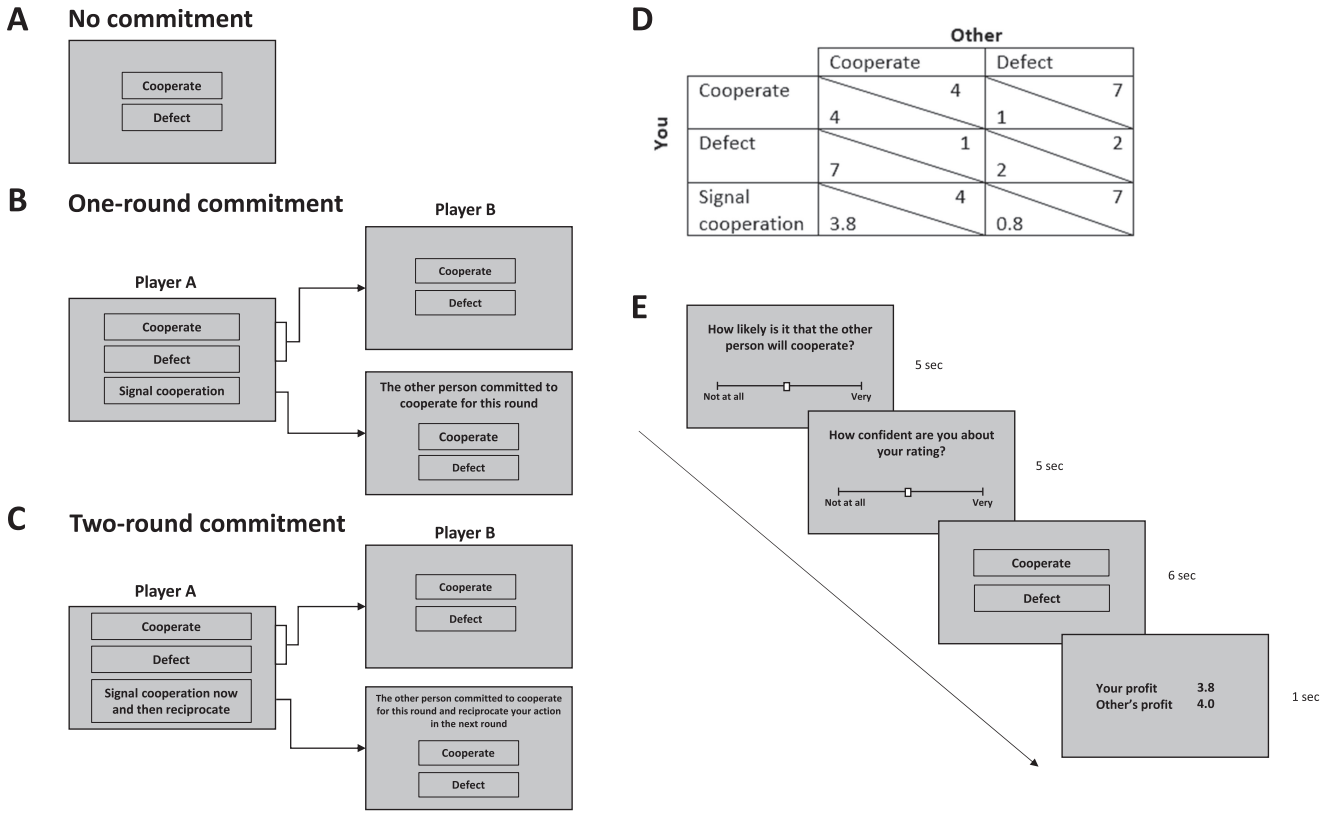
In the *one-round commitment* condition, in addition to the cooperation and defection options participants could choose to costly signal cooperation for the current round. In this condition, the game was played sequentially, with Player A deciding before Player B. If Player A decided to cooperate or to defect, then Player B had to make a choice without knowing Player A's decision (so Players A and B virtually decided simultaneously), as in the no-commitment condition. However, if Player A selected to signal cooperation to Player B, then the game became sequential and Player B was informed that Player A committed to unilateral cooperation for the current trial ("The other person committed to cooperate for this round"), such that Player B decided to cooperate or defect based on the knowledge about Player A's commitment (Figure 1B). When Player A decided to commit, 0.2 experimental coins were subtracted from the current trial's profit (i.e., they would obtain 3.8 instead of four coins in case of mutual cooperation and 0.8 coins in the case that Player B defected). The profits of Player B remained unaffected (Figure 1D). Thus, if the commitment option was chosen, our task paradigm became similar to sequential versions of the prisoner's dilemma game (Ahn et al., 2007; Clark & Sefton, 2001), though in our paradigm signaling was costlier than unsignaled cooperation (which is not the case in standard sequential prisoner's dilemma games). Note that in the current paradigm, commitment decisions were always honest (i.e., participants could not decide to defect after having signaled cooperation), such that in our study the costs an individual was willing to pay for commitment did not serve as an indicator of the individual's honesty, as assumed by costly signaling theory (Laidre & Johnstone, 2013; Smith & Bliege Bird, 2005).

In the *two-round commitment condition*, instead of committing to cooperation for only one round, Player A could commit to a tit-for-tat strategy over two rounds. This allowed us to compare the effectiveness of one-round commitments to cooperation with commitments to a tit-for-tat strategy, which represents the most prominent and efficient strategy in the prisoner's dilemma game (Axelrod & Hamilton, 1981). In addition to the options "Cooperate" and "Defect," participants could select "Signal cooperation now and then reciprocate." By selecting this option, participants signaled to Player B that they would cooperate in the current trial t and would automatically reciprocate Player B's action in the following trial ($t + 1$). If Player A selected this option, the signaling cost (0.2 coins) was subtracted only on trial t, and Player B was informed about Player A's commitment in the current round before making a choice (Figure 1C). As before, if Player A selected either to cooperate or defect, the trial continued as in the no-commitment condition. Note that when considering only the two rounds after a commitment decision it might appear more beneficial for Player B to defect (expected payoff: $7 + 2 = 9$) than to cooperate (expected payoff: $4 + 4 = 8$). However, if the interaction continued after these two rounds (e.g., if commitment was chosen at the start of a block), then Player B was better off when cooperating instead of defecting after Player A's commitment choices. In both the one-round and the two-round commitment conditions, Player B was aware of the availability of commitment options to Player A.

To measure the influence of participants' predictions of the co-player's behavior as well as how confidence in these predictions affects commitment decisions, participants answered two questions before each decision in all three conditions. Participants first had to indicate the likelihood that the other player would cooperate on a continuous rating scale within 5 s. Because predicting others' future actions requires estimating the probability that the interaction partner will cooperate or defect, we asked participants to give their prediction on a continuous scale, expressed as the likelihood (probability) that the other will cooperate (Miettinen et al., 2020; Trautmann & van de Kuilen, 2015). We note that in paradigms where binary predictions were required (cooperate vs. defect), participants are forced to map the predicted likelihood of the other's behavior onto the predefined binary answer format (Brosig, 2002; Sparks et al., 2016). In contrast, our quantitative assessment of predictions provided us with a more refined measure of participants' beliefs about the other's next action. Finally, they rated their confidence in their prediction on a scale ranging from *not at all (confident)* to *very*. The confidence rating allowed participants to express a second-order judgment on the correctness of their first-order rating (Lebreton et al., 2015). Participants provided ratings on visual analog scales, on which they had to slide the marker via mouse click (Figure 1E).

The task was divided into blocks of eight rounds against the same co-player. After each block, participants were randomly matched to

Figure 1
Experimental Paradigm and Task Conditions



Note. (A) In the *No commitment* condition, participants made their decisions simultaneously between the two options to cooperate or defect, without knowing the co-player's decision. (B) In the *One-round commitment* condition, Player A had the additional option to signal her decision to cooperate to Player B. If Player A selected to signal cooperation, Player B was notified and could use this information to make her decision. (C) In the *Two-round commitment* condition, Player A had the additional option to commit to a tit-for-tat strategy over two rounds. If Player A selected to commit to the strategy, Player B was informed of this decision. (D) Payoff matrix. If both players cooperated, they both received four experimental coins each, whereas if both defected, they received two experimental coins each. If one player cooperated and the other defected, the cooperator received one coin and the free-rider seven coins. If Player A selected to signal her choice to cooperate, she incurred a cost of 0.2 coins. (E) Trial timeline. At the start of the trial, participants were asked to predict the co-player's behavior (5 s) and subsequently to rate their confidence in their prediction (5 s). Following the confidence rating, participants made their decision for the corresponding trial (6 s) and received feedback on the joint outcome, that is, their own and the co-player's profit for the current trial (1 s).

another anonymous participant in the room. The task condition as well as the role of participants as Player A or B changed randomly after each block. We administered four blocks for each condition, resulting in a total of 96 rounds.

Procedure

The experiment was conducted in groups of eight participants, and participants were not allowed to communicate with each other during the experiment in order to ensure the anonymity of the decisions. After reading the instructions, they performed a practice block of the prisoner's dilemma task for 20 trials, where they practiced all experimental conditions. Following the practice, block participants performed the main prisoner's dilemma task. After the prisoner's dilemma, participants completed an individual decision-making task, four short questionnaires including the Social Value Orientation scale (Murphy et al., 2011), the perspective-taking scale of the Questionnaire for Cognitive and Affective Empathy (Reniers et al., 2011), a short risk attitude questionnaire, where they selected between different gambles and certain monetary alternatives, and finally a short demographic questionnaire. None of these measures correlated with signaling rates (all $|\text{tau}| < 0.015$, all $ps > .394$). The demographic questionnaire included questions for age, field of study, mother tongue (free response), and questions for level of education (multiple choice; high school/bachelor/master/PhD) and gender (multiple choice; male/female/other). At the end of the experiment, the accumulated payoff in the prisoner's dilemma game was exchanged to euros at a rate of 100 coins = €1 and paid out to the participants in addition to their standard compensation.

Statistical Analysis

Statistical analyses were performed using R, Version 4.0.3 (R Core Team, 2020). For all analyses, alpha threshold was set to 5%.

As preregistered, we used generalized linear mixed models (GLMMs) to test whether commitment choices were predicted by fixed-effects predictors for Prediction, Confidence, Condition (one-round commitment, two-round commitment, with the reference category set to the former) as well as the interaction between these variables using the lme4 package (Bates et al., 2015). We *z*-standardized all continuous variables on the individual level to control for individual differences in prediction as well as metacognitive bias in the different conditions (Galvin et al., 2003; Maniscalco & Lau, 2012; Soutschek et al., 2021). This ensures that the relationship between commitment choices and confidence is not confounded by individual differences in metacognitive bias (i.e., that some individuals are generally more confident than others). We further modeled participant-specific random intercepts and random slopes for all within-subject predictors. To measure cooperation rates, we calculated mutual cooperation pairs (Signal–Cooperate or Cooperate–Cooperate) for each participant and condition. To test whether cooperation rates differed between conditions, we used the nonparametric Friedman test, as the dependent variable was not normally distributed (Shapiro–Wilk normality test, $p < .05$). Finally, we computed correlations with the Kendall rank correlation coefficient (Kendall's τ).

Results

Participants decided for the signaling option with a probability of 34% in the one-round commitment and with 20% in the two-round commitment conditions, suggesting that participants indeed used the commitment options when available. Both in the one-round and the two-round conditions, participants preferred to use the commitment options at the start of the interaction block over cooperation or defection decisions, and the frequency of commitment choices dropped over the course of a block ($\beta = -0.828$, $z = -8.073$, $p < .001$, Cohen's $d = 1.16$), as indicated in a generalized linear model where we regressed the decision to commit on predictors for trial, as illustrated in Figure 2A and Figures S1 and S3 in the online supplemental materials. This suggests that the commitment option is used to signal one's willingness to cooperate with the co-player, which, while still being used by some participants throughout the interaction block, is less needed once mutual cooperation has been established (note that in the two-round commitment condition, Player A's choice was predetermined only if Player A had committed in the previous round). With lack of evidence on the interaction partner's prior behavior at the beginning of a block, confidence in one's prediction was lower at the start compared to the end of block (mean confidence at trial 1 = 36.19 and at trial 8 = 59.87), as indicated by a main effect of trial ($\beta = 7.65$), $t(47) = 4.634$, $p < .001$, Cohen's $d = 0.66$, in a linear mixed model where we regressed confidence ratings on predictors for trial (Figure 2B and Figures S2 and S4 in the online supplemental materials).

We first tested the hypothesis that the possibility for costly commitment enhances cooperation rates by assessing whether the percentage of mutual cooperation pairs (Signal–Cooperate or Cooperate–Cooperate) significantly differed between the three task conditions with the Friedman test. Our analysis revealed no significant difference in cooperation rates between conditions, $\chi^2_F(2) = 4.25$, $p = .120$, indicating that the introduction of costly commitment options per se is not enough to significantly promote cooperation compared to the standard prisoner's dilemma. To further explore whether the availability of the two commitment options contributed to differences in earnings from the task, we tested whether accumulated payoffs differed among the three conditions. Again, the Friedman test revealed no significant difference among the three conditions, $\chi^2_F(2) = 4.01$, $p = .135$. Thus, across all participants, the availability of commitment options did not increase overall cooperation rates.

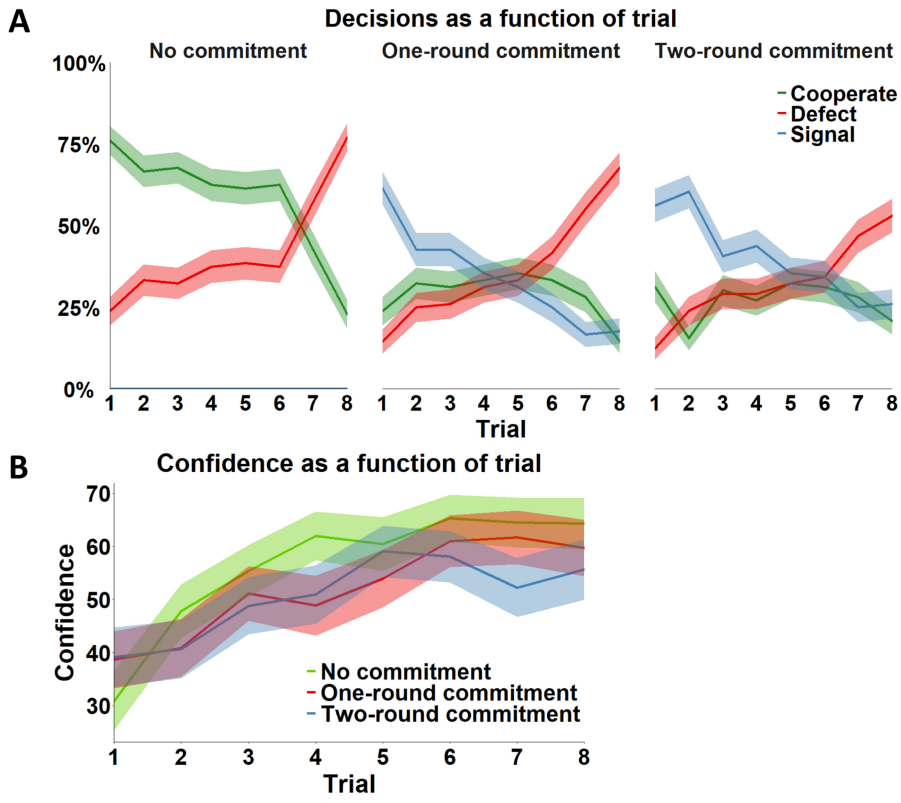
Lastly, we tested whether the absence of commitment in the two commitment conditions might have been interpreted by Player B as potential defection of Player A. For this, we compared Player B's defection rates between the commitment and no commitment conditions when Player A did not previously signal cooperation. If the absence of signaling led to a different interpretation of nonsignaled cooperative decisions, this should be reflected by a higher defection rate in the commitment conditions when Player A did not signal. We found no significant difference between defection rates in the commitment and no-commitment conditions in the absence of signaling cooperation ($Z = -1.37$, $p = .170$, $r = .19$). Therefore, there was no evidence that the availability of the signaling option changed Player B's interpretation of Player A's cooperation versus defection decisions, which might be explained by the assumption that Player B was aware that Player A's likelihood of commitment was not stable but dropped over the course of an interaction.

Next, we tested the hypothesis that participants make commitment choices particularly when their confidence in predicting the co-player's intentions is low. We regressed binary coded choices to commit (1 = commitment, 0 = no commitment) on fixed-effects predictors of Condition, Prediction, Confidence, and their interaction terms. To account for the possibility that the link between prediction and commitment follows a quadratic (participants might commit particularly when they cannot clearly predict the co-player's intentions) rather than a linear relationship, we also included the quadratic term of Prediction as regressor to our model. In the one-round commitment condition, the probability of choosing to commit increased with lower confidence ($\beta = -0.40$, $z = -2.020$, $p = .043$, Cohen's $d = 0.29$; Figure 3A and Figures S5 and S7 in the online supplemental materials), and with higher predicted probability that Player B would cooperate ($\beta = 0.73$, $z = 2.423$, $p = .015$, Cohen's $d = 0.34$). This supports our hypothesis that participants decide to signal their cooperativeness to the other predominantly when they are uncertain about the other's intentions. Finally, while the Confidence × Prediction interaction showed only a trend-level effect ($\beta = 0.31$, $z = 1.812$, $p = .069$, Cohen's $d = 0.26$), the quadratic term for Prediction was significant, with $\beta = -0.35$, $z = -2.689$, $p = .007$, Cohen's $d = 0.38$, suggesting an inverted U-shaped relationship between the predicted likelihood of cooperation and the willingness to commit (i.e., participants choose the commitment option predominantly when they cannot clearly predict whether the other will cooperate or defect). No other effect reached significance, suggesting that there were no significant differences between the one-round and two-round commitment conditions (Table S1 in the online supplemental materials). Note that the significant main effect of confidence was robust to excluding the quadratic term for prediction ($p = .017$), as well as both the linear and quadratic effect of prediction ($p = .033$) from the GLMM.

To further exclude the possibility that the effect of confidence on deciding to signal cooperation could be the result of the correlations between both variables and trial number, we regressed commitment choices on predictors for Confidence, Prediction, and Trial number. While the effect of Trial number was significant ($\beta = -0.76$,

Figure 2

Overview of Decisions and Confidence Within an Interaction Block



Note. (A) Percentage rates of each decision as a function of trial in Experiment 1. Across conditions, cooperation rates dropped toward the end of the interaction block. Both at the one-round and two-round conditions, more participants were selected to start the interaction block by signaling their commitment to the other person. Error bands represent the standard error of the mean. (B) Mean confidence ratings in each condition as a function of trial in Experiment 1. Across conditions, confidence ratings increased toward the end of the interaction block. Error bands represent the standard error of the mean. See the online article for the color version of this figure.

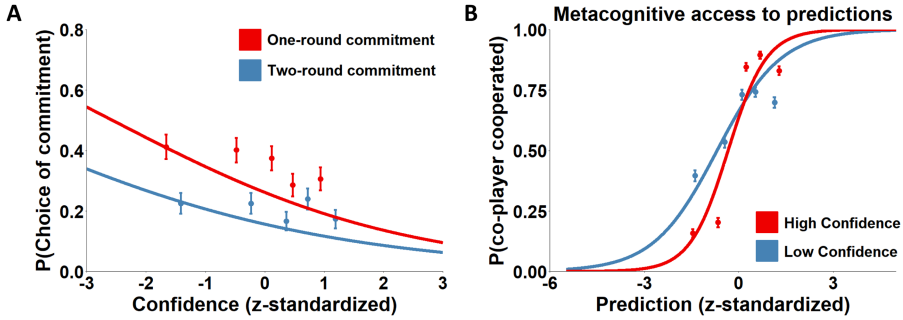
$z = -8.900$, $p < .001$, Cohen's $d = 1.28$), the effect of Confidence on decisions to signal commitment remained significant ($\beta = -0.21$, $z = -2.031$, $p = .042$, Cohen's $d = 0.29$), as did the effect of Prediction ($\beta = 0.40$, $z = 2.519$, $p = .011$, Cohen's $d = 0.36$) and their interaction ($\beta = 0.25$, $z = 2.422$, $p = .015$, Cohen's $d = 0.34$; Table S2 in the online supplemental materials).

Moreover, one might argue that the observed link between confidence and commitment is confounded with participants' ability to adjust their decisions based on the other's behavior. For example, discrepancies between the other's predicted and actual choices might lower participants' confidence in their ability to predict the other's behavior and motivate commitment choices. From this perspective, the correlation between confidence and commitment would be caused by the influence of prediction errors on these two variables. To exclude the possibility that the observed results stem from pure reinforcement learning effects, we first analyzed the data only from the first trial of each interaction block, which should be free from any learning effects. While the effect of confidence was not significant across conditions ($\beta = -0.29$, $z = -0.957$, $p = .338$, Cohen's $d = 0.13$), the correlation between confidence and commitment was significant in the one-round commitment condition ($\beta = -4.32$, $z = -2.297$, $p = .021$, Cohen's $d = 0.33$). We note, however, that the reliability of this analysis might be questionable, as restricting our sample to the first trial of each block resulted in a dataset of only six trials per participant and condition. Therefore, to account for both learning effects and the co-player's past behavior, we ran a further analysis that controlled for the influence of prediction errors by regressing commitment choices on predictors for Confidence, Prediction, and Prediction Error. While the effect of prediction error on commitment choices was not significant ($\beta = -0.00$, $z = -0.004$, $p = .996$, Cohen's $d = 0.00$), Prediction again had a positive effect on commitment ($\beta = 0.55$, $z = 2.844$, $p = .004$, Cohen's $d = 0.41$), whereas Confidence showed a significant negative effect ($\beta = -0.27$, $z = -2.075$, $p = .037$, Cohen's $d = 0.29$; Table S3 in the online supplemental materials). Thus, our results do not support the alternative explanation according to which the correlation between confidence and commitment is driven by reinforcement learning effects.

The observed link between subjective confidence and willingness to commit raises the question as to whether participants can reliably

Figure 3

Results From Experiment 1



Note. (A) Probability of commitment as a function of confidence in mentalizing in Experiment 1. Participants were more likely to select to commit to cooperation when subjective confidence in their prediction was low. This effect was significant in both the one-round and the two-round commitment conditions. Regression lines are based on the group-level fixed effect parameter estimates from the regression models. Raw data are binned and overlaid over regression lines for reference. Error bars represent standard error of the mean. (B) Metacognitive access to mentalizing performance. Participants possessed metacognitive access to their mentalizing performance such that they assigned lower confidence to trials where their prediction was wrong and assigned higher confidence to the correct predictions. For illustration purposes, we split our sample into low- and high-confidence trials (median split). Regression lines are based on the group-level fixed effect parameter estimates from the regression models. Raw data are binned and overlaid over regression lines for reference. Error bars represent the standard error of the mean. See the online article for the color version of this figure.

track the accuracy of their mentalizing processes (metacognitive accuracy). To test this, we regressed the co-player's decisions (1 = signal cooperation/cooperate, 0 = defect) on fixed-effects predictors of Prediction, Confidence, and the interaction term. A positive Confidence × Prediction interaction would indicate that participants can track (and overtly report) the strength of the correlation between their predictions and the other's actual choices (De Martino et al., 2013). In fact, we found a significant Confidence × Prediction interaction ($\beta = 0.29$, $z = 3.897$, $p < .001$, Cohen's $d = 0.56$), suggesting that participants possess metacognitive access to the accuracy of their mentalizing processes (Figure 3B and Table S4 in the online supplemental materials). In other words, they can reliably track and report when a prediction is likely to be correct, and when it is not.

Discussion

Experiment 1 provides empirical evidence for a link between confidence in mentalizing processes and costly commitment in social interactions: participants committed to costly signaling particularly when they had low confidence in their prediction of the other's intention. Interestingly, we found no significant difference between the one-round and the two-round commitment conditions. Low confidence motivated the use of signaling both commitments to cooperation and commitment to a strategy. It is possible that in the case of the two-round commitment, where reciprocation of defection is among the potential outcomes, costly commitment could be associated with norm enforcement rather than solely with signaling, an interpretation that could extend to both conditions in the absence of a significant difference on the decision-driving factors. Noteworthy, however, the structure of the tit-for-tat commitment was such that Player B could defect during a two-round commitment from Player A, earning more coins than if Player B cooperated (nine instead of eight coins). Thus, two-round commitment might have also been used as a tool to reduce the risk of defection rather than a norm enforcement device. Moreover, to the best of our knowledge, norm enforcement has not been previously associated with confidence in mentalizing, and therefore the evidence that low confidence could motivate (commitment to) norm enforcement is limited.

We further showed that commitment was preferred if participants could not clearly predict the co-player's intentions and if participants predicted the other to cooperate rather than to defect. Finally, we showed that participants have metacognitive access to the accuracy of their mentalizing process. While these findings provide evidence for the importance of metacognitive access to mentalizing processes in social interactions, the observed link between confidence and costly signaling was only correlative in nature. The goal of Experiment 2, therefore, was to assess whether experimentally manipulating subjective confidence could affect the willingness to commit.

Experiment 2

In Experiment 2, we replicated our findings from Experiment 1 and furthermore established a causal link between confidence in mentalizing and costly commitment. To experimentally lower subjective confidence, we manipulated the amount of feedback participants received in the prisoner's dilemma game. In line with previous literature, we expected that lowering the amount of evidence on the other's behavior through omitting feedback reduces confidence in predictions of the other's intentions (Boldt et al., 2017; Desender et al., 2019). Based on the findings of Experiment 1, we hypothesized that lower confidence in mentalizing in conditions

with rare feedback should increase the willingness to costly signal cooperation.

Materials and Method

Participants

Thirty-two volunteers were recruited through the participant pool of the Melessa lab at the Ludwig Maximilian University Munich, Germany. The sample size was based on the same power analysis as for Experiment 1, though due to lab closings during the COVID-19 pandemic in Fall 2020, a further experimental session with four participants had to be canceled. Data from one participant were dropped due to familiarity with the paradigm, resulting in a final sample size of 31 participants (15 female; 16 male; $M_{age} =$ 23.5 years, $SD = 4.16$). All participants gave informed written consent and received compensation for their participation, which consisted of a base fee of €12 and an additional bonus depending on their choices in the task.

Stimuli and Task Design

Prisoner's Dilemma. Participants played the prisoner's dilemma task in the one-round commitment condition in the same way as in Experiment 1 (Figure 1B, 1D, and 1E). To experimentally reduce participants' confidence in their predictions of the other's behavior, we manipulated the amount of feedback participants received at the end of each trial in three conditions. In the 100% feedback condition, participants received feedback about both their and the other's payoffs (e.g., "Your Profit: 4, Other's Profit: 4") after every decision, as in Experiment 1. In the 50% feedback condition, participants received feedback about their own and the other's outcome after 50% of all trials. In the 25% feedback condition, we further reduced the feedback rate to 25%. In the 50% and 25% conditions, feedback was presented randomly after both players' decisions with the respective probabilities. If the trial was a no-feedback trial, the participants simply moved to the next trial after making their decision and without knowing what the other person had decided.

Study Design and Procedure

The experiment was conducted in groups of four participants. Again, all interactions were anonymous. After reading detailed instructions for the task, participants performed a practice block for 20 trials, where they practiced all conditions and player types. Following the practice block participants performed the main prisoner's dilemma task with a total of 96 trials (32 trials per condition, divided into blocks of eight trials). After each block, participants were randomly and anonymously matched to a different participant in the room. The feedback condition as well as the role of participants as Player A and Player B changed randomly every eight rounds.

Statistical Analysis

As in Experiment 1, we used generalized linear models to test the relationship between commitment choices and confidence in mentalizing. We regressed commitment choices on fixed-effects predictors of Prediction, Confidence, Feedback (100%, 50%, 25%, with the reference category set to 100%), and the interaction terms. All continuous variables were z-standardized separately for each participant and each condition to control for individual differences in prediction as well as metacognitive bias (Galvin et al., 2003; Maniscalco & Lau, 2012; Soutschek et al., 2021). All models included participant-specific random intercepts and within-subject fixed effect predictors as random slopes. Because predictions and confidence ratings were not normally distributed (Shapiro–Wilk normality test, $p < .05$), we used the Friedman test to assess differences in prediction and confidence across feedback conditions. For pairwise comparisons, we used the paired Wilcoxon signed-rank test, where p values were adjusted using the Bonferroni correction method.

Results

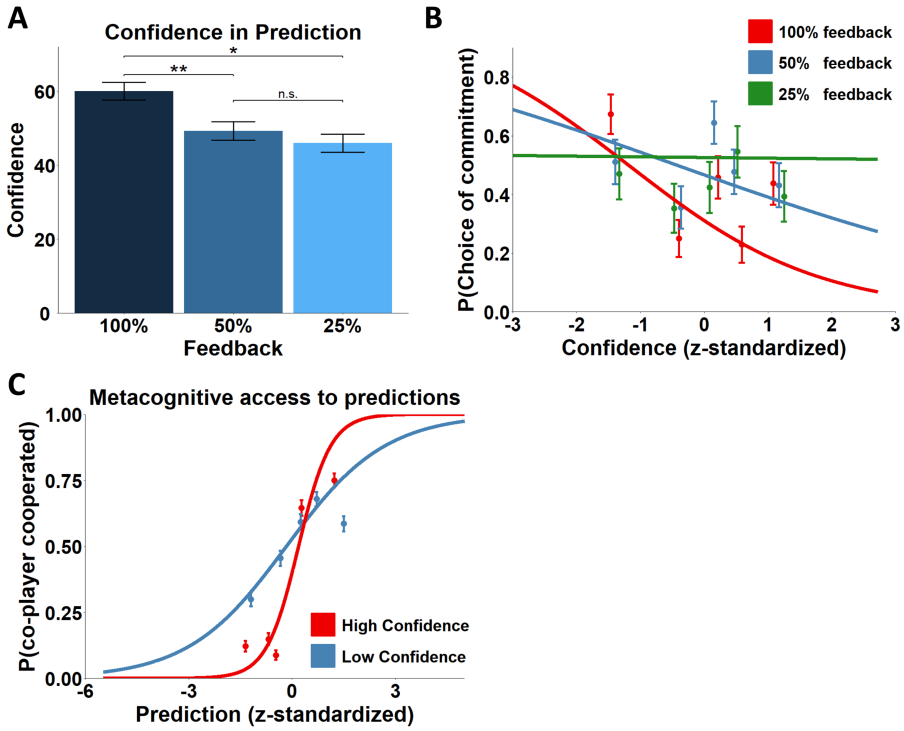
First, we aimed to replicate our previous findings by testing whether lower confidence in predicting the co-player's intentions is linked to more commitment choices in the 100% feedback condition. For this purpose, we regressed binary choices to signal cooperation in the current round (1 = signal cooperation, 0 = cooperate or defect) in the 100% condition on fixed-effects predictors for Prediction, Confidence, and the interaction effect. The main effect of Confidence ($\beta = -0.87$, $z = -3.180$, $p = .001$, Cohen's $d =$ 0.62) replicated the finding of Experiment 1 that lower confidence in predictions is linked with more commitment.

Next, we performed a sanity check and tested whether participants indeed have lower confidence in their predictions in the 50% and 25% feedback conditions compared to the 100% feedback condition. For this, we calculated mean confidence ratings per participant in each condition and tested whether there was a difference in confidence ratings among the three conditions using the Friedman test. We found significant differences in confidence between feedback conditions, $\chi^2_F(2) = 9.51$, $p = .008$ (Figure 4A). Pairwise comparisons revealed that confidence ratings were significantly lower in the 50% ($M = 49.31$) and 25% ($M = 46.33$) feedback conditions compared to the 100% ($M = 60.39$) feedback condition ($Z = -2.80$, $p_{\text{Bonferroni-corrected}} = .005$, $r = .52$ and $Z = -1.96$, $p_{\text{Bonferroni-corrected}} = .049$, $r = .36$, respectively). Finally, there was no significant difference in confidence ratings between the 50% and the 25% feedback conditions ($Z = 0$, $p_{\text{Bonferroni-corrected}} = 1$, $r = 0$), suggesting that while less reliable evidence can reduce the confidence in one's mentalizing process relative to the 100% feedback condition, there might exist a threshold where confidence in mentalizing cannot further be reduced.

Next, in order to test our hypothesis that less feedback is associated with a higher probability of commitment, we regressed binary choices to signal cooperation for the current round (1 = signal cooperation, 0 = cooperate) on fixed-effects predictors for Feedback, Prediction, Confidence, and the interaction effects. Note that defection rates were significantly increased in the 25% feedback condition ($M = 64.65\%$) compared with the 100% feedback ($M = 46.12\%$) and 50% feedback ($M = 51.72\%$) conditions (both $p < .05$), which may confound influences of confidence on commitment choices given that commitment is more likely if the other player is expected to cooperate rather than defect. To control for this confound, we focused on the willingness to commit relative to the preference for (uncommitted) cooperation. This analysis revealed that compared to the 100% feedback condition, participants were more likely to signal their cooperation both in the 50% ($\beta = 0.66$, $z =$

Figure 4

Results From Experiment 2



Note. (A) Mean confidence levels across the three feedback conditions. Compared to 100% feedback, participants were overall less confident in their predictions both in the 50% and in the 25% feedback conditions (Bonferroni-corrected). There was no significant difference in confidence ratings between the 50% and 25% feedback conditions. Error bars represent the standard error of the mean. (B) Effect of confidence on choice of commitment. Probability of commitment as a function of confidence in mentalizing for the three feedback conditions. Participants were more likely to select to commit to cooperation in the low feedback conditions compared to the 100% feedback condition. The effect of confidence on commitment in the 100% feedback condition, such that lower confidence predicted higher probability of commitment replicated the findings from Experiment 1. Regression lines are based on the group-level fixed effect parameter estimates from the regression models. Raw data are binned and overlaid over regression lines for reference. Error bars represent the standard error of the mean. (C) Metacognitive access to mentalizing performance in Experiment 2. Participants demonstrated metacognitive access to their mentalizing performance such that they rated the correct predictions of their co-player's behavior with higher confidence. As with Experiment 1, we split our dataset into high- and low-confidence trials (median split) for illustration purposes. Regression lines are based on the group-level fixed effect parameter estimates from the regression models. Raw data are binned and overlaid over regression lines for reference. Error bars represent the standard error of the mean. See the online article for the color version of this figure.

2.178, $p = .029$, Cohen's $d = 0.42$) and in the 25% feedback condition ($\beta = 0.89$, $z = 2.897$, $p = .003$, Cohen's $d = 0.56$). Moreover, the significant main effect of Confidence ($\beta = -0.66$, $z = -2.775$, $p = .005$, Cohen's $d = 0.54$; Table S5 in the online supplemental materials) supports our previous findings that lower confidence is associated with more commitment choices. The significant Feedback (25%) × Confidence interaction ($\beta = 0.65$, $z = 2.024$, $p = .042$, Cohen's $d = 0.39$) suggests that participants in the 25% feedback condition might have a higher threshold of confidence for signaling cooperation (Figure 4B and Figures S6 and S8 in the online supplemental materials). A separate GLMM indicated no significant difference between the 50% and 25% feedback conditions ($\beta = 0.23$, $z = 0.761$, $p = .446$, Cohen's $d = 0.14$). These results support our hypothesis that less feedback (and thus higher uncertainty in predicting the other's intentions) increases the preference for costly signaling cooperation over unsignaled cooperation.

Additionally, as in Experiment 1, we tested the effect of Trial number on decisions to signal commitment to cooperation by regressing commitment choices on predictors for Confidence, Prediction, and Trial number. As before, the effect of Trial number was significant ($\beta = -0.83$, $z = -7.622$, $p < .001$, Cohen's $d = 1.41$), and also the main effect of Confidence remained significant ($\beta = -0.47$, $z = -3.099$, $p = .001$, Cohen's $d = 0.57$; Table S6 in the online supplemental materials). Furthermore, we again controlled for reinforcement learning effects, first by analyzing data from the first trial of each interaction block and second by regressing

commitment choices on predictors for Confidence, Prediction, and Prediction Error. Again, despite its limited reliability, the analysis of choices in the first trial of each interaction block revealed a trend-level effect of confidence across conditions ($\beta = -0.45$, $z = -1.729$, $p = .083$, Cohen's $d = 0.32$). Finally, we found no effect of prediction errors on commitment choice ($\beta = -0.38$, $z = -1.401$, $p = .161$, Cohen's $d = 0.26$), whereas we replicated the significant negative effect of Confidence ($\beta = -0.99$, $z = -3.709$, $p < .001$, Cohen's $d = 0.70$; Table S7 in the online supplemental materials). Also in Experiment 2, there was thus no evidence for reinforcement learning effects on commitment choices.

Finally, we sought to replicate our finding that participants have metacognitive awareness of their mentalizing accuracy during social interactions. When regressing the other's choices on Confidence, Prediction, and the interaction term, we again found a significant Confidence × Prediction interaction ($\beta = 0.41$, $z = 3.611$, $p < .001$, Cohen's $d = 0.65$; Table S8 in the online supplemental materials), replicating our finding from Experiment 1 that participants can metacognitively access and evaluate their mentalizing performance (Figure 4C).

Discussion

Experiment 2 replicated our finding that confidence in predicting the interaction partner's intentions is linked to the selection of costly commitment. By experimentally reducing participants' confidence in their predictions, we provided evidence that uncertainty about the other's actions motivates commitment decisions and further replicated our previous finding that participants possess metacognitive access to their mentalizing processes.

General Discussion

Mentalizing is an important skill in social interactions but cannot always be performed with high accuracy (Wu et al., 2020). Theoretical accounts suggest that internal representations of the accuracy of mentalizing processes are an important precondition for successful social interactions (Han et al., 2015). Here, we put these theoretical models on an empirical ground and advance the field of social interactions by providing evidence for a crucial role of metacognitive processes in social decision-making. First, in two independent experiments, we show that participants possess metacognitive access to the accuracy of their mentalizing processes, as agents can track and report the reliability of their predictions of others' intentions. Second, this metacognitive insight guides agents' choices to commit as they resort to bear the costs of signaling their willingness to cooperate, particularly when they have low confidence in their reading of the other's intentions. Finally, we show that there is a causal link, and not just a correlation, between uncertainty in predicting others' actions and the willingness to use costly signals of cooperation.

Our finding that low confidence comes with a higher probability to commit is in line with a computational model showing that strategic decision-making is more efficient when agents incorporate their subjective confidence into the decision process (Han et al., 2015). While previous studies have traditionally focused on the importance of mentalizing in social interactions (Brosig, 2002; Fischbacher et al., 2001), here we provide empirical evidence that, in addition to predicting the other's behavior, subjective confidence in these predictions (and thus metacognitive insight in one's mentalizing performance) plays an equally crucial role in strategic interactions. According to recent accounts, mentalizing is essential in social interactions to translate external information (such as others' past behaviors) into predictions of others' future actions and to update one's beliefs about these intentions (Molapour et al., 2021; Thornton & Tamir, 2021; Wu et al., 2020). Subjective confidence in mentalizing, therefore, reflects the perceived reliability of these predictions and may play a significant role in evaluating and updating them (Molapour et al., 2021). In strategic social interactions, it is thus essential not only to have an estimate of what the other will do, but also an internal representation of how reliable this estimate is.

For confidence to efficiently guide strategic decision-making, subjective confidence needs to be related to the objective accuracy of mentalizing processes. In fact, our results reveal that participants possess metacognitive access to the reliability of intention recognition processes, providing new evidence for the relevance of metacognition in social decision-making (Frith, 2012; Wu et al., 2020). The findings are in line with what is shown for individual decisions, that is, that metacognition influences decision-making through optimizing the use of commitment devices (Soutschek et al., 2021; Soutschek & Tobler, 2020). Here, we showed that metacognition is an integral component of strategic interactions and speculate that better metacognitive abilities can potentially contribute to optimizing social decision processes, such that decision-makers, through evaluating their mentalizing process, can spare unnecessary costs and yield more benefits from the interaction.

Commitment in strategic games has been shown to be preferred in coordination games such as Battle-of-the-Sexes or Hawk-Dove game, facilitating the achievement of the signaler's preferred outcomes and leading to long-term benefits. Binding commitment to a particular course of action provided an advantage to the signalers in contexts forcing the co-player to adhere to the equilibrium set by the signaler. Moreover, participants were willing to pay for having the possibility to commit, and the bid they placed was typically matched with the expected value of committing (Barclay, 2017). Furthermore, it seems that signaling choices are preferred in games close to the Prisoner's Dilemma, Stag Hunt, or Harmony Games, where the signal can serve to increase trust between the parties and promote long-term cooperation (Barclay et al., 2021).

Importantly, signaling is not always beneficial and highly depends on the level of conflict present in the paradigm. For example, no commitment is needed when conflict is absent, as there is no substantial benefit in deviating from coordinating, whereas in high conflict contexts committing can even be harmful, allowing the other party to adjust their decisions to the signaler's disadvantage. Recent research has shown that costly commitment is preferred in situations involving moderate levels of conflict and a clear benefit of committing (Barclay, 2017). Similarly, our results can generalize to such contexts, but further research will be needed to characterize how low confidence influences commitment in different conflict contexts and how the cost a signaler is willing to pay depends on the relative uncertainty in their predictions.

Our findings provide evidence that subjective confidence in mentalizing performance is a crucial driver of social decision-making. Given that mentalizing is an integral part of social interactions (Wu et al., 2020), we expect our results to generalize to other domains of social decision making including interactions where agents have the opportunity to signal their intentions to others.

However, how our results generalize to social decision-making, in general, needs to be tested with different paradigms with varying levels of conflict and signaling options. While our results may be contingent on metacognitive ability, which varies on the individual level (Fleming et al., 2010; Rouault et al., 2018), we have no reason to believe that our findings depend on other characteristics of our sample. Across both experiments, we tested male and female subjects, and the majority of participants had either completed high school or bachelor studies of varying fields. Thus, we believe our findings will be reproducible with participants from similar populations, though it needs to be determined whether our findings hold also for clinical populations or different age groups.

An extension to clinical populations could be particularly interesting, as mentalizing deficits come as a hallmark symptom in several psychiatric disorders (e.g., autism spectrum disorder, borderline personality disorder, schizophrenia; Chung et al., 2014; Fonagy & Bateman, 2008) and recent evidence also suggest a crucial role of overconfidence in schizophrenia (Köther et al., 2012). Further considering the role of metacognition in social cognition may deepen our understanding of these disorders and inspire new therapeutic approaches focusing on this neglected but indispensable component of social interactions.

In conclusion, our research sheds new light on the intricacies of social interactions and how they are influenced by our own self-awareness. We have discovered that our confidence in our ability to understand others' intentions plays a crucial role in our decision-making, particularly when it comes to signaling our willingness to cooperate. By considering not only our ability to read others' minds but also our own perception of how reliable this process is, we can gain a more nuanced understanding of the complex dynamics at play in our social lives.

References

Ahn, T.-K., Lee, M., Ruttan, L., & Walker, J. (2007). Asymmetric payoffs in simultaneous and sequential prisoner's dilemma games. *Public Choice*, *132*(3–4), 353–366. https://doi.org/10.1007/s11127-007-9158-9

Arieli, I., Babichenko, Y., & Tennenholtz, M. (2017). Sequential commitment games. *Games and Economic Behavior*, *105*, 297–315. https://doi.org/10.1016/j.geb.2017.08.009

Axelrod, R., & Hamilton, W. D. J. S. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396. https://doi.org/10.1126/science.7466396

Balliet, D. (2010). Communication and cooperation in social dilemmas: A meta-analytic review. *Journal of Conflict Resolution*, *54*(1), 39–57. https://doi.org/10.1177/0022002709352443

Barclay, P. (2017). Bidding to commit: An experimental test of the benefits of commitment under moderate degrees of conflict. *Evolutionary Psychology*, *15*(1), Article 147470491769074. https://doi.org/10.1177/1474704917690740

Barclay, P., Bliege Bird, R., Roberts, G., & Számadó, S. (2021). Cooperating to show that you care: Costly helping as an honest signal of fitness interdependence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1838), Article 20200292. https://doi.org/10.1098/rstb.2020.0292

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Boldt, A., de Gardelle, V., & Yeung, N. (2017). The impact of evidence reliability on sensitivity and bias in decision confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(8), 1520–1531. https://doi.org/10.1037/xhp0000404

Brosig, J. (2002). Identifying cooperative behavior: Some experimental results in a prisoner's dilemma game. *Journal of Economic Behavior & Organization*, *47*(3), 275–290. https://doi.org/10.1016/S0167-2681(01)00211-6

Buskens, V., & Royakkers, L. (2002). Commitments: A game-theoretic and logical perspective. *Cognitive Science Quarterly*, *2*(3/4), 448–467.

Chung, Y. S., Barch, D., & Strube, M. (2014). A meta-analysis of mentalizing impairments in adults with schizophrenia and autism spectrum disorder. *Schizophrenia Bulletin*, *40*(3), 602–616. https://doi.org/10.1093/schbul/sbt048

Clark, K., & Sefton, M. (2001). The sequential prisoner's dilemma: Evidence on reciprocation. *The Economic Journal*, *111*(468), 51–68. https://doi.org/10.1111/1468-0297.00588

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110. https://doi.org/10.1038/nn.3279

Deroy, O., Spence, C., & Noppeney, U. (2016). Metacognition in multisensory perception. *Trends in Cognitive Sciences*, *20*(10), 736–747. https://doi.org/10.1016/j.tics.2016.08.006

Desender, K., Boldt, A., Verguts, T., & Donner, T. H. (2019). Confidence predicts speed-accuracy tradeoff for subsequent decisions. *eLife*, *8*, e43499. https://doi.org/10.7554/eLife.43499

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., Fehr, E., Stephan, K. E., & Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Computational Biology*, *10*(9), Article e1003810. https://doi.org/10.1371/journal.pcbi.1003810

Elitzur, R., & Gavious, A. (2003). Contracting, signaling, and moral hazard: A model of entrepreneurs, 'angels,' and venture capitalists. *Journal of Business Venturing*, *18*(6), 709–725. https://doi.org/10.1016/S0883-9026(03)00027-2

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178. https://doi.org/10.1007/s10683-006-9159-4

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404. https://doi.org/10.1016/S0165-1765(01)00394-9

Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *The Journal of Neuroscience*, *32*(18), 6117–6125. https://doi.org/10.1523/JNEUROSCI.6489-11.2012

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541–1543. https://doi.org/10.1126/science.1191883

Fonagy, P., & Bateman, A. (2008). The development of borderline personality disorder—A mentalizing model. *Journal of Personality Disorders*, *22*(1), 4–21. https://doi.org/10.1521/pedi.2008.22.1.4

Frank, R. H., Gilovich, T., & Regan, D. T. (1993). Does studying economics inhibit cooperation? *Journal of Economic Perspectives*, *7*(2), 159–171. https://doi.org/10.1257/jep.7.2.159

Frith, C. D. (2012). The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2213–2223. https://doi.org/10.1098/rstb.2012.0123

Gächter, S. (2006). *Conditional cooperation: Behavioral regularities from the lab and the field and their policy implications* (CeDEx Discussion Paper No. 2006-03). University of Nottingham.

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, *10*(4), 843–876. https://doi.org/10.3758/BF03196546

Han, T. A., Pereira, L. M., & Santos, F. C. (2011). Intention recognition promotes the emergence of cooperation. *Adaptive Behavior*, *19*(4), 264–279. https://doi.org/10.1177/1059712311410896

Han, T. A., Santos, F. C., Lenaerts, T., & Pereira, L. M. (2015). Synergy between intention recognition and commitments in cooperation dilemmas. *Scientific Reports*, *5*(1), Article 9312. https://doi.org/10.1038/srep09312

Harrington, J. E., & Zhao, W. (2012). Signaling and tacit collusion in an infinitely repeated prisoners' dilemma. *Mathematical Social Sciences*, *64*(3), 277–289. https://doi.org/10.1016/j.mathsocsci.2012.05.005

Kapetaniou, G. E., Deroy, O., & Soutschek, A. (2022). *Social metacognition drives willingness to commit* [Data set]. https://osf.io/eq3f2/?view_only=7ff216408a9c4e24a114f9c53c0ac268

Keser, C., & Van Winden, F. (2000). Conditional cooperation and voluntary contributions to public goods. *The Scandinavian Journal of Economics*, *102*(1), 23–39. https://doi.org/10.1111/1467-9442.00182

Köther, U., Veckenstedt, R., Vitzthum, F., Roesch-Ely, D., Pfueller, U., Scheu, F., & Moritz, S. (2012). "Don't give me that look"—Overconfidence in false mental state perception in schizophrenia. *Psychiatry Research*, *196*(1), 1–8. https://doi.org/10.1016/j.psychres.2012.03.004

Krach, S., Blümel, I., Marjoram, D., Lataster, T., Krabbendam, L., Weber, J., van Os, J., & Kircher, T. (2009). Are women better mindreaders? Sex differences in neural correlates of mentalizing detected with functional MRI. *BMC Neuroscience*, *10*(1), Article 9. https://doi.org/10.1186/1471-2202-10-9

Laidre, M. E., & Johnstone, R. A. (2013). Animal signals. *Current Biology*, *23*(18), R829–R833. https://doi.org/10.1016/j.cub.2013.07.070

Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience*, *18*(8), 1159–1167. https://doi.org/10.1038/nn.4064

Locey, M. L., & Rachlin, H. (2012). Commitment and self-control in a prisoner's dilemma game. *Journal of the Experimental Analysis of Behavior*, *98*(1), 89–103. https://doi.org/10.1901/jeab.2012.98-89

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Miettinen, T., Kosfeld, M., Fehr, E., & Weibull, J. (2020). Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions. *Journal of Economic Behavior & Organization*, *173*, 1–25. https://doi.org/10.1016/j.jebo.2020.02.018

Molapour, T., Hagan, C. C., Silston, B., Wu, H., Ramstead, M., Friston, K., & Mobbs, D. (2021). Seven computations of the social brain. *Social Cognitive and Affective Neuroscience*, *16*(8), 745–760. https://doi.org/10.1093/scan/nsab024

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. J. (2011). Measuring social value orientation. *Judgment and Decision Making*, *6*(8), 771–781. https://doi.org/10.1017/S1930297500004204

Nakamura, M., & Ohtsuki, H. (2016). Optimal decision rules in repeated games where players infer an opponent's mind via simplified belief calculation. *Games*, *7*(3), Article 19. https://doi.org/10.3390/g7030019

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Reniers, R. L., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A questionnaire of cognitive and affective empathy. *Journal of Personality Assessment*, *93*(1), 84–95. https://doi.org/10.1080/00223891.2010.528484

Renou, L. (2009). Commitment games. *Games and Economic Behavior*, *66*(1), 488–505. https://doi.org/10.1016/j.geb.2008.05.001

Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human metacognition across domains: Insights from individual differences and neuroimaging. *Personality Neuroscience*, *1*, e17. https://doi.org/10.1017/pen.2018.16

Smith, E. A., & Bliege Bird, R. (2005). Costly signalling and cooperative behaviour. In H. Gintis, S. Bowles, R. Boyd & E. Fehr (Eds.), *Moral sentiment and material interests: The foundations of cooperation in economic life* (pp. 115–148). MIT Press.

Soutschek, A., Moisa, M., Ruff, C. C., & Tobler, P. N. (2021). Frontopolar theta oscillations link metacognition with prospective decision making. *Nature Communications*, *12*(1), Article 3943. https://doi.org/10.1038/s41467-021-24197-3

Soutschek, A., & Tobler, P. N. (2020). Know your weaknesses: Sophisticated impulsiveness motivates voluntary self-restrictions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(9), 1611–1623. https://doi.org/10.1037/xlm0000833

Sparks, A., Burleigh, T., & Barclay, P. (2016). We can see inside: Accurate prediction of prisoner's dilemma decisions in announced games following a face-to-face interaction. *Evolution and Human Behavior*, *37*(3), 210–216. https://doi.org/10.1016/j.evolhumbehav.2015.11.003

Thornton, M. A., & Tamir, D. I. (2021). The organization of social knowledge is tuned for prediction. In: M. Gilead & K. N. Ochsner (Eds.) *The neural basis of mentalizing* (pp. 283–297). Springer.

Trautmann, S. T., & van de Kuilen, G. (2015). Belief elicitation: A horse race among truth serums. *The Economic Journal*, *125*(589), 2116–2135. https://doi.org/10.1111/ecoj.12160

Vaccaro, A. G., & Fleming, S. M. (2018). Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. *Brain and Neuroscience Advances*, *2*, Article 239821281881059. https://doi.org/10.1177/2398212818810591

Valk, S. L., Bernhardt, B. C., Böckler, A., Kanske, P., & Singer, T. (2016). Substrates of metacognition on perception and metacognition on higher-order cognition relate to different subsystems of the mentalizing network. *Human Brain Mapping*, *37*(10), 3388–3399. https://doi.org/10.1002/hbm.23247

Wu, H., Liu, X., Hagan, C. C., & Mobbs, D. (2020). Mentalizing during social InterAction: A four component model. *Cortex*, *126*, 242–252. https://doi.org/10.1016/j.cortex.2019.12.031

Yamaguchi, M., Smith, A., & Ohtsubo, Y. (2015). Commitment signals in friendship and romantic relationships. *Evolution and Human Behavior*, *36*(6), 467–474. https://doi.org/10.1016/j.evolhumbehav.2015.05.002

Yoshida, W., Dolan, R. J., & Friston, K. J. (2008). Game theory of mind. *PLoS Computational Biology*, *4*(12), Article e1000254. https://doi.org/10.1371/journal.pcbi.1000254

Received March 2, 2022
Revision received February 27, 2023
Accepted March 2, 2023 ■