PREPRINT

Adaptive curiosity about metacognitive ability

Samuel Recht[1], Canqi Li[2*], Yifan Yang[3*] & Kaiki Chiu[4]

[1] Department of Experimental Psychology, University of Oxford, Oxford, UK
[2] Department of Cognitive, Linguistic & Psychological Sciences, Brown University, Providence, USA
[3] Department of Chinese Language and Literature, Tsinghua University, Beijing, China
[4] Department of Cognitive Science, Barnard College, Columbia University, New York, USA

* These authors contributed equally

Corresponding author: SR, samuel.recht@gmail.com

Abstract

Metacognition provides control and oversight to the process of acquiring and using knowledge. Efficient metacognition is essential to many aspects of daily life, from healthcare to finance and education. Across three experiments, we found a specific form of curiosity in humans about the quality of their own metacognition, using a novel approach that dissociates cognitive from metacognitive information searches. Observers displayed a strategic balance in their curiosity, alternating between a focus on cognitive accuracy and metacognitive performance. This adaptive curiosity was modulated by an internal evaluation of metacognition, leading to increased feedback requests when metacognition was likely to be inaccurate. These results show that individuals are inherently curious about their metacognitive abilities and can compare cognitive and metacognitive precision to fine-tune performance monitoring. We propose that this newly observed curiosity may reflect humans' focus on meta-learning, or 'learning to learn', a promising avenue in the study of minds and machines.

Keywords: curiosity, metacognition, perception, self-regulation

Introduction

Metacognition – humans' ability to reflect on their own cognition – has been shown to play an important role in regulating behaviour, providing oversight and control to the process of acquiring knowledge and understanding. From infancy to adulthood and across neurotypes, metacognition is often viewed as a critical factor in human's adaptability, and its implication in fostering learning has been recently extended to the study of artificial systems (Lake & Baroni, 2023; Wang et al., 2021; Cox, 2005).

Metacognition is usually defined through two main components: monitoring of first-order thoughts, perceptions, feelings, and memories, and the active control over these mental operations to improve their efficiency in dealing with the environment (Fleming et al., 2012; Yeung & Summerfield, 2012). A key dimension of metacognition could be found in confidence judgement, which determines our certainty about our decisions and sensory experiences (Mamassian, 2016). Confidence shapes how we seek information, prioritise tasks, set goals, and communicate (e.g., Aguilar-Lleyda & de Gardelle, 2021; Pescetelli et al., 2021; Desender et al., 2018; Bahrami et al., 2012). How well confidence judgments reflect reality is often referred to as 'metacognitive ability' and has been shown to vary across individuals and situations. Despite being often reliable, metacognitive accuracy is prone to fluctuation, particularly early in development and in certain clinical populations.

Many aspects of daily life can be damaged by metacognitive failures, from healthcare to finance, policy making and education. For instance, generalised anxiety disorder and major depressive disorder are often linked with exaggerated perceptions of negative outcomes (e.g., Fu et al., 2012), while in bipolar disorder, compromised metacognitive monitoring can delay the recognition of mood episodes, potentially worsening the condition (Van Camp et al., 2019). In financial decision-making, overconfidence can lead to increased risk-taking and sensation-seeking behaviours (Stotz & von Nitzsch, 2005; Grinblatt & Keloharju, 2009), and has been tied to the destruction of company value (Ahmed & Duellman, 2013). Overconfidence is also a feature of increased risk taking in the sharing of sensitive information over the internet (Sawaya et al., 2017). Within the political spectrum, individuals holding radical beliefs have been shown to exhibit lower metacognitive accuracy, suggesting a role of metacognitive failure in extreme polarisation (Rollwage et al., 2018). Finally, ineffective metacognition among students has been described as one of the main obstacles to their academic growth (Bransford et al., 2000). Hence, the negative impact of metacognitive failures on society are significant enough that efforts should be made to better understand and address them.

Improving metacognitive evaluation in this variety of contexts and personal experiences requires nuanced and individualised feedback strategies (e.g., debiasing in finance, Kaustia & Perttula, 2012). As confidence level plays a significant role in regulating exploration during value-based decision (Boldt et al., 2019) and perception (Desender et al., 2018), enhancing metacognitive skills can lead to better decision-making through improved control and monitoring. For example, guiding patients specifically to identify and address their metacognitive shortcomings has shown promise in psychotherapy (e.g., Fisher & Wells, 2008; Lysaker et al., 2018). Metacognitive interventions have also been found to improve

resilience against misinformation (Salovich & Rapp, 2021), and to significantly foster education (Hattie, 2008). Yet, whether humans can effectively probe and improve their own metacognition without explicit guidance or training remains a largely open question.

Recent work proposes that metacognitive monitoring could involve even further levels of self-evaluation, opening the possibility that individuals may recognise metacognitive failures and remediate them in a self-supervised manner. In a recent study (Recht et al., 2022), we found evidence of 'meta-metacognitive' ability in human vision (e.g., ‘I know that I know that I saw’, or ‘Type-3’ cognition) without the need for training or explicit feedback. Subsequent experiments using different research paradigms have further corroborated these findings (Zheng et al., 2023; Sherman & Seth, 2023). This ability to identify errors in one's own metacognition could notably prove crucial for developing more effective learning strategies (Händel & Fritzsche, 2016; Buratti & Allwood, 2015; Hattie, 2009). Although humans can assess their metacognition, this does not guarantee they will do so willingly without explicit instructions or clear benefits.

A crucial step, therefore, involves examining how people value metacognitive information: are individuals intrinsically interested in understanding their own thought processes? If so, how do they weigh the significance of insights about their metacognition against more direct cognitive information? One challenge in addressing these questions lies in the closely intertwined relationship between perception and metacognition. It is generally understood that metacognitive evidence is at least partially built upon perceptual decision evidence, thereby linking searches for information on one aspect inevitably to the other. In this study, we introduce a novel approach that separates these two forms of evidence, enabling us to explore how decision evidence influences the preference for seeking information about metacognitive abilities.

We propose that the strength of perceptual evidence presented to an observer may influence their curiosity about their metacognitive abilities rather than on their perceptual performance. To test our hypothesis, we designed a paradigm where participants made two perceptual judgments (or first-order decisions), followed by a metacognitive judgement to select the better of the two initial decisions (Mamassian, 2020). A last stage involved giving the option to seek feedback on the accuracy of either perception or metacognition. Crucially, many participants preferred feedback about their metacognitive decision, despite only perception feedback offering direct information about their gain in a given trial. To provide a mechanistic description of this behaviour, we propose an observer model in which evidence is collected and evaluated at different cognitive levels, and then compared between levels to orient curiosity.

Methods

Participants

In Experiment 1, 20 naïve adults were recruited from the Prolific platform, averaging 31 ± 7 years (M ± SD; 7 females) and compensated at £8.5/hour. Experiment 2 included 10 students from the University of Oxford, reimbursed with course credits, and an additional 16 participants, recruited via word-of-mouth, totaling 26 participants (25 ± 11 years, 15 females).

In Experiment 3, 62 adults from the Prolific platform (31 ± 8 years, 29 females), participated for £8.5/hour. Exp 1 and Exp 3 took approximately 20 min to complete, with the 50% best performing participants receiving a bonus of £0.6. Exp 2 took 50 min to complete, and no bonus was paid. All participants across the experiments reported normal or corrected vision, no known neurological issues, no psychoactive medication use, fluency in English, and had an approval rate equal or above 97% on Prolific for Exp 1 and 3. All procedures were approved by the University of Oxford's Research Ethics Committee. Exp 1 was exploratory without formal power analysis. Sample sizes for Exp 2 and 3 were based on Exp 1's effect size, using a Bayesian stopping rule. Supplementary Material provides further details. In Exp 2, 1 out of the initial 26 participants were excluded due to failed attention checks. Similarly, in Experiment 3, 1 out of the 62 participants was excluded for the same reason, and another one for not submitting any data (because of a technical issue). In total, data from 105 participants were used for further analysis. Both Exp 2 and 3 were pre-registered on AsPredicted.org (details provided in the Supplementary Material).

Experimental protocol

Each trial started with a central fixation cross (0.6°) presented on a light grey background for 100ms in Exp 1, 800ms in Exp 2 and 500ms in Exp 3 (inter-trial interval or 'ITI'). Following the ITI, a circle of 9.53° – assuming a distance of 60 cm from the screen – and colour RGB [187, 187, 187] was displayed at fixation, with 20 equally spaced 2.86° visual gratings composing a grid (spatial frequency = 0.05 cycle/degree, Gaussian window). Upon obtaining informed consent, participants were provided with detailed instructions on how to perform the task. Participants were encouraged to give unspeeded responses for all of their responses. To familiarise participants with the task and the format of the trials, they were then presented with four demo trials that did only include the perceptual discrimination. These demo trials included visual feedback on the accuracy of their responses. Then, following the demo trials, the actual experiment started: on each trial, participants were asked to first observe the gratings within the circle for 2 seconds and estimate their average orientation (**Figure 1**). Then, participants were asked to reproduce that average orientation with their mouse cursor. Each response was converted into points depending on accuracy (from 0 points corresponding to a 90° error, up to 100 points, corresponding to a 0° error). The experiment was displayed in full-screen mode. Exiting full-screen mode during the course of the experiment was automatically reversed in the following trial.

Every two trials, participants were asked to choose one of the two previous responses to keep for reward (confidence 2-alternative forced choice, or 2AFC), knowing that only the points from the selected response will be added to their total score. Importantly, following the confidence 2AFC, participants were presented with a 4-alternative forced choice (4AFC) - with the options listed in random order - to decide what type of feedback they would like to receive: no feedback, feedback on first response, feedback on second response, or feedback on final choice (i.e., metacognitive judgement). If they decided to choose the first or second response, they were provided with the value of their response in points. If they opted for the final choice, they were informed whether the response they kept was the best one (but without its value in points). Finally, the 'no feedback' option led to no information. Following their click, the button turned to the requested information (or remained empty for the 'no

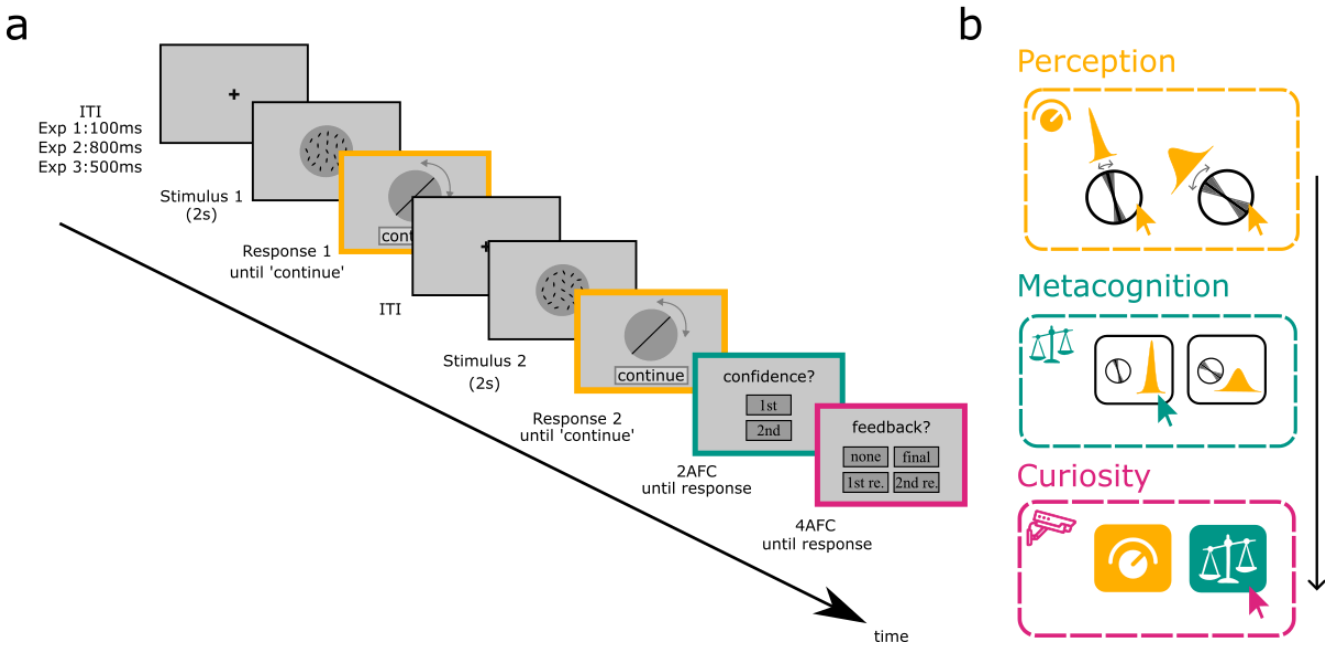

Figure 1. Paradigm. **(a)** In a perceptual decision task, participants were presented with a grid of oriented gratings for 2s, and had to reproduce the average orientation at the end of the trial (i.e., perceptual decision). After two trials, they had to make a confidence judgement (i.e., metacognitive decision) by selecting which of the two previous trials was the best. Finally, they were given the option (i.e., curiosity decision) to seek feedback either on their perceptual or metacognitive accuracy. They also had the option to not request any feedback at all. The inter-trial interval (ITI) differs, as well as the number of participants/trials and the payoff structure (Exp 1 and 3 involved payment and bonuses, but not Exp 2). **(b)** From top to bottom panels: Each panel represents a decision order. In the first 'Perception' panel, the observer makes perceptual decisions, with varying degrees of precision (represented by the width of the distribution around the correct angle). The second 'Metacognition' panel represents the selection of the best trial in the pair of perceptual tasks, with the aim of selecting the one with greater precision. The bottom 'Curiosity' panel represents the choice in feedback requests, where participants have to decide between perceptual or metacognitive feedback (see Figure 2a for the distribution of responses).

feedback' option) for 1.2s, followed by an inter-trial interval that varied between experiments (100, 800, or 500 ms). Every 20 trials, participants were prompted to provide three subjective distinct global estimates of their performance within the block using a rating scale. The order of these questions were randomised. One more question involved an attentional check to make sure participants were paying attention to the task. Except for the attentional check, we did not analyse these ratings in the present study. Following the rating scales, participants were provided with a 15-second break to rest. During this break, they were presented with the number of points earned in the previous 20 trials, along with the total points earned throughout the experiment. Importantly, we manipulated the relative difficulty level of trials within a pair by altering the variance of the stimulus grid (and therefore the difficulty of reproduction), in four distinct conditions: easy-easy ('E-E'), where the variance for the stimuli in the two trials was equated and low, easy-hard ('E-H') where the variance was larger for the second trial in the pair, hard-easy ('H-E') and hard-hard ('H-H'). All three experiments had the same characteristics, at the exception of the ITI, the number of trials / participants, and the

payoff structure. Specifically, Exp 1 involved 80 trials with 20 participants, Exp 2 involved 200 trials and 25 participants and Exp 3, 80 trials and 60 participants.

Results

The experimental design and main adaptive curiosity hypothesis were preregistered for Exp 2 and 3. Exp 1 was an exploratory study. In the following sections, we report all statistical tests we have used for our hypotheses and explicitly state when a test was preregistered. We used linear mixed-effects models (see Supplementary Material for details) and report ΔAIC as the difference between the model without the tested effect and the model with the effect (a negative value is evidence against the tested effect).

Perception (Type-1 decision)

Participants' perceptual accuracy was quantified using the absolute difference between the reported orientation and the true orientation (i.e., the mean orientation of the gratings in the grid). This error was also converted into points (max. of 100 points). On average, participants did well (M ± SD; Exp 1: 81.88 ± 8.14; Exp 2: 81.95 ± 8.79; Exp 3: 77.55 ± 11.80). The error in degrees is an index of precision. We expected lower error on average for the 'easy' trial pairs ('E-E'), medium average error in the mixed pairs ('E-H' or 'H-E'), and greater average error for the 'hard' pairs ('H-H'). **Figure 2b**, top panel illustrates the relative difference in error between conditions. We used a linear mixed-effects regression to test the association between condition and trial order (i.e., whether it was the first versus second trial in a pair) as predictor(s), and perceptual error (outcome). Across all three experiments, we found that condition was a significant predictor of error (all χ^2s > 47.50, *ps* < .001, ΔAICs > 48.50, Table S1). The main effect of trial order was not significant (all χ^2s < 1.51, *p*s > .21, ΔAICs < 1.70, Table S2), but we found an interaction between trial order and condition (all χ^2s > 90.82, *ps* < .001, ΔAICs > 91.43, Table S3). The interaction is expected given the existence of mixed pairs ('E-H' vs. 'H-E'). Such results confirmed successful manipulation of difficulty level across conditions, which effectively influenced participants' perceptual performance.

Metacognition (Type-2 decision)

In a second analysis, we quantified participants' metacognitive accuracy. During the metacognitive (or Type-2) judgments, participants had to select the best of their two previous responses to keep for later reward. Here, metacognitive accuracy is the probability of correctly selecting the trial with lower error. In a set of *t*-tests, we compared the average metacognitive accuracy to chance-level (50%). As expected, participants showed above-chance metacognitive accuracy, being able to pick out the more precise response in all experiments (all *p*s < .001, BF_{10}s > 388.70, Table S4). As per perceptual accuracy, we also expected condition to be a significant factor for metacognitive performance (**Figure 2b**, middle panel). Among the four conditions, we expected that it is more difficult to distinguish between similar pairs ('E-E' and 'H-H'), while it is easier to do so in mixed pairs ('E-H' and 'H-E'). Consequently, mixed trial pairs should lead to greater metacognitive accuracy compared to similar ones. For all three experiments, we found a main effect of condition in predicting metacognitive accuracy (all χ^2s > 8.90, *p*s < .035, ΔAICs > 2.90, Table S5). Such

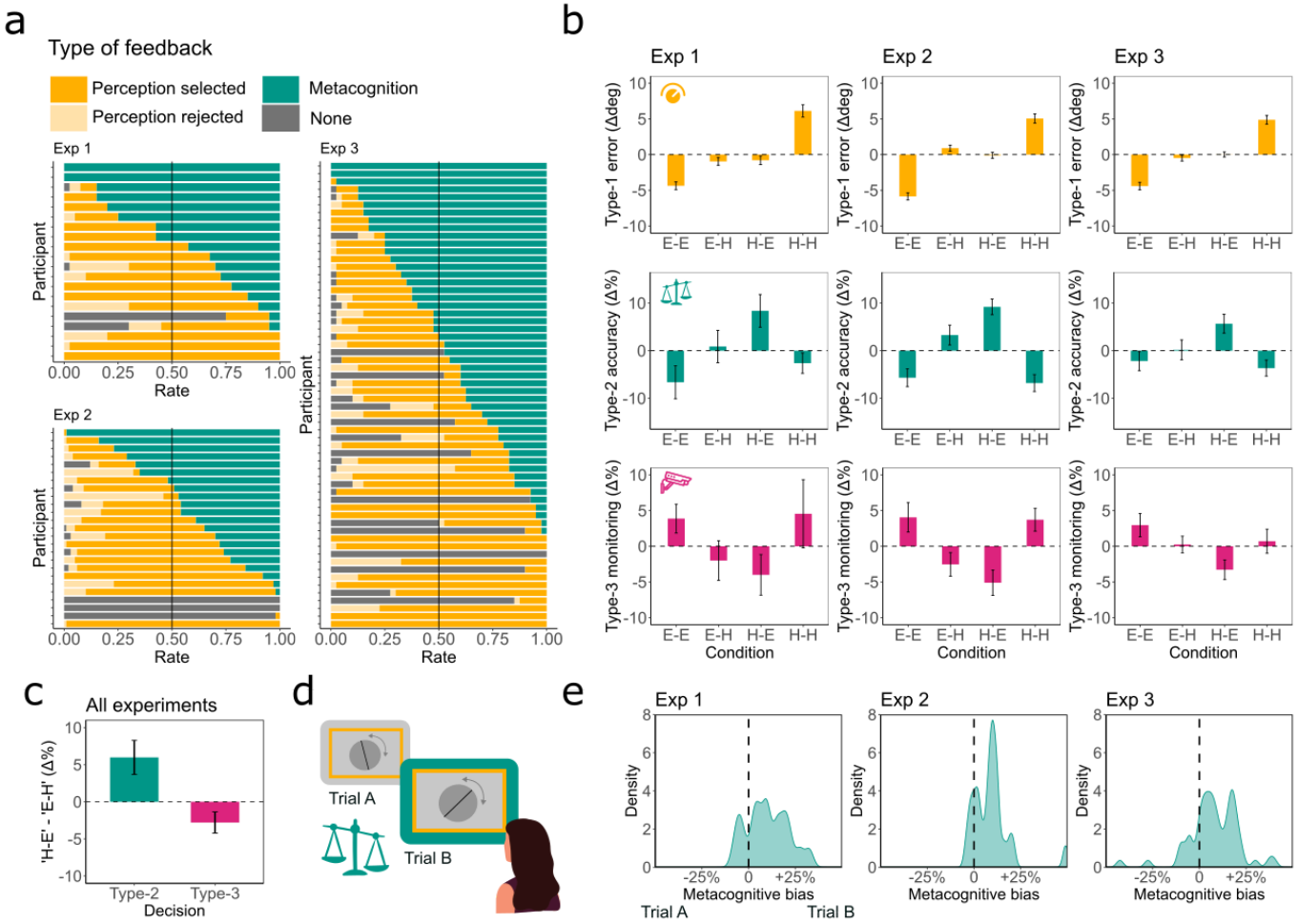

Figure 2. Behaviour. **(a)** The chart displays the distribution of feedback requests per participant for each experiment. In the chart, the yellow bars represent the proportion of feedback requests related to the quality of the perceptual decision (Type-1) for both the selected and rejected (low alpha) trials. Notably, the majority of Type-1 feedback requests focused on the selected trials. The green bars represent the proportion of feedback requests related to the quality of the metacognitive decision (Type-2). The grey bars indicate the proportion of instances where participants did not request any specific type of feedback. **(b)** The figure illustrates the average accuracy for perception (yellow), metacognition (green), and the proportion of times participants requested Type-2 (metacognitive) feedback instead of Type-1 (perceptual) feedback across different conditions (purple, Easy-Easy, Easy-Hard, Hard-Easy and Hard-Hard). On the y-axis, the values represent the average difference in absolute error in degrees (Δdeg) or the average difference in percent (Δ%), normalised per participant. **(c)** The difference between Hard-Easy and Easy-Hard for metacognitive accuracy (Type-2) and curiosity (Type-3) request rates, pooled across experiments, showing higher metacognitive accuracy for the Hard-Easy compared to Easy-Hard, despite both having the same metacognitive difficulty. **(d) and (e)** The difference observed in (c) can be understood when considering the overall metacognitive 'recency' bias towards the second trial in the pair. Figure (e) plots the distribution of average bias across participants, the majority of participants showed a bias towards the second trial (i.e., were more likely to select the second trial). Error bars represent the standard error of the mean (SEM).

results suggested that participants were above chance in their metacognitive judgments and that our condition manipulation also affected their metacognitive performance. An

unexpected finding was the tendency for metacognitive accuracy to be greater in the 'H-E' compared to the 'E-H' condition, despite the two conditions involving similar difficulty levels for metacognition (**Figure 2c**). An exploratory analysis (over data pooled across experiments) confirmed a significant accuracy difference ($t(104) = 2.60$, $p = .01$, $BF_{10} = 2.61$). This effect was probably driven by the metacognitive bias towards the second trial in the pair observed in all experiments (**Figure 2d** and **2e**; all ts > 4.06, ps < .001 and BF_{10}s > 52.60, Table S6): when the condition conformed to the bias, metacognitive performance increased.

Curiosity (Type-3 decision)

Our main aim was to investigate how participants arbitrate feedback-seeking about the quality of their perceptual and metacognitive judgments. A first prediction was that despite only perceptual performance feedback directly informing on the actual points earned, participants would still show an interest in metacognitive feedback, even at the cost of not getting perceptual feedback at all. Considering the proportion of metacognitive compared to perceptual feedback requests (filtering out the 'no feedback' trials - two participants who selected this option in every 4AFC in Exp 2 were excluded), we found very strong evidence for this curiosity for metacognitive ability in all experiments. In one-sample t-tests, the curiosity for metacognitive feedback was systematic (all Ms > .40, ps < .001, BF_{10}s > 820.40, Table S7). Participants were, on average, more curious about the quality of their metacognition rather than perception more than 40% of the time. **Figure 2a** illustrates the notable inter-individual variability for such an appetite: some participants were mostly interested in their perceptual accuracy (in yellow in the figure), while other participants were mostly interested in their metacognitive accuracy (in green). Many participants also showed alternations over trials. Comparing the curiosity for perceptual accuracy to that for metacognitive accuracy using t-tests, we found that the request rate for metacognitive feedback was not significantly below 50% overall (ps > 0.13, BF_{10}s < 0.65, Table S8). The similar average curiosity for the two types of feedback suggested that at the population level, participants showed a strong appetite for metacognitive feedback.

Strategic shifts in curiosity between perception and metacognition

Similar to metacognitive accuracy, we further hypothesised that the curiosity for metacognitive feedback is also a function of the difficulty posed by the metacognitive task (i.e., how difficult it is to select the better trial in a pair). Participants would request feedback on their metacognitive choice more often when it is challenging (between similar trials), but less so when it is easy (between mixed trials). Using a linear mixed-effects logistic regression (preregistered for Exp 2-3), we found the condition to significantly predict the rate of metacognitive feedback requests in two out of the three experiments. In both Exp 2 and 3, the models with condition as a predictor were significantly better than the null models (Exp 2: $\chi^2(3) = 21.05$, $p < .001$, $\Delta AIC = 15.05$; Exp 3: $\chi^2 = 8.34$, $p = .04$, $\Delta AIC = 2.34$), but not in Exp 1 ($\chi^2(3) = 3.74$, $p = .29$, $\Delta AIC = -2.26$). The fact that Exp 2 (involving a larger number of trials than Exp 1) and Exp 3 (involving a larger number of participants than Exp 1) showed the main effect of condition suggested that the lack of effect was probably due to a lower statistical power. **Figure 2b**, lower panel, illustrates the relative Type-3 curiosity as a function

of condition, once normalised per participant (to account for individual's average curiosity, for readability).

Finally, while the condition effect suggested an effect of metacognitive difficulty on curiosity, our last prediction pertained to the effect of metacognitive accuracy itself. If participants are able to evaluate the quality of their metacognition (a signature of meta-metacognitive ability), they should seek corresponding feedback on metacognition particularly when their metacognitive choice is wrong. In a linear mixed-effects logistic regression (preregistered for Exp 2-3), we found metacognitive accuracy to predict curiosity in two out of three experiments. In both Exp 1 and 3, the models including metacognitive accuracy as a predictor were significantly better than the null models (Exp 1: $\chi^2(1) = 6.89$, $p = .01$, $\Delta AIC = 4.89$; Exp 3: $\chi^2(1) = 5.42$, $p = .02$, $\Delta AIC = 3.42$). The model for Exp 2 did not offer significant improvement over the null model ($\chi^2(1) = 0.15$, $p = .69$, $\Delta AIC = -1.85$). We found no significant interaction between metacognitive accuracy and condition in Exp 1 and 2 (all χ^2s < 8.40, ps > .20, ΔAICs < -3.65). In contrast, in Exp 3, the model with interaction improved the fit significantly ($\chi^2(3) = 11.25$, $p = .01$, $\Delta AIC = 5.25$, Table S9). Consistent with the results, one-sample *t*-tests (preregistered for Exp 2-3) also showed significant differences in request rates of metacognitive feedback between wrong and correct metacognitive choices in Exp 1 and 3 (only considering participants having data points in each category; all ps = .01, BF_{10}s > 2.99, Table S10), but not in Exp 2 ($p = .78$, $BF_{10} = 0.24$). Following our exploratory analysis on metacognitive bias (**Figure 2c-e**), we also tested a potential difference in curiosity between 'H-E' and 'E-H' pairs. Despite a qualitative trend in all experiments (**Figure 2b-c**), the difference remained at significance threshold ($t(98) = -1.95$, $p = .05$, $BF_{10} = 0.69$).

Bayesian observer model

To understand in more detail the nature of the observed curiosity tradeoff, we developed a Bayesian observer model (**Figure 3a**). In the model, for each pair of trials, the observer first estimates the orientation of the stimulus (perceptual decisions, Figure 3a left panel), and then compares the precision in each decision, selecting the one with greater precision (metacognitive response, Figure 3a middle panel). For curiosity, the observer then compares perceptual evidence to metacognitive evidence to request feedback (Figure 3a right panel). This strategy would allow the observer to fine-tune information intake depending on the context and curiosity traits, considering that both perceptual and metacognitive evidence are subjectively valuable. The valuation being always in perceptual evidence units, this approach provides a form of common currency across perception, metacognition and curiosity judgments. Drawing from earlier research (e.g., Van den Berg et al., 2012, 2014), we use Fisher information (J) to measure perceptual precision and quantify the evidence used for metacognitive and curiosity judgements. We hypothesise that perception fluctuates from trial to trial, leading to variable precision in reports. To model these fluctuations, we opted for a Gamma distribution (two parameters, see Supplementary Material), a model used in both the working memory (Schneegans et al., 2020; Van den Berg, 2012) and metacognition literature (Geurts et al. 2022; Recht et al. 2021). As shown on Figure 3a, the nature of perceptual evidence has a significant impact on predicted metacognitive accuracy and curiosity across precision levels. When most of the change in perception reflects a change in the scale

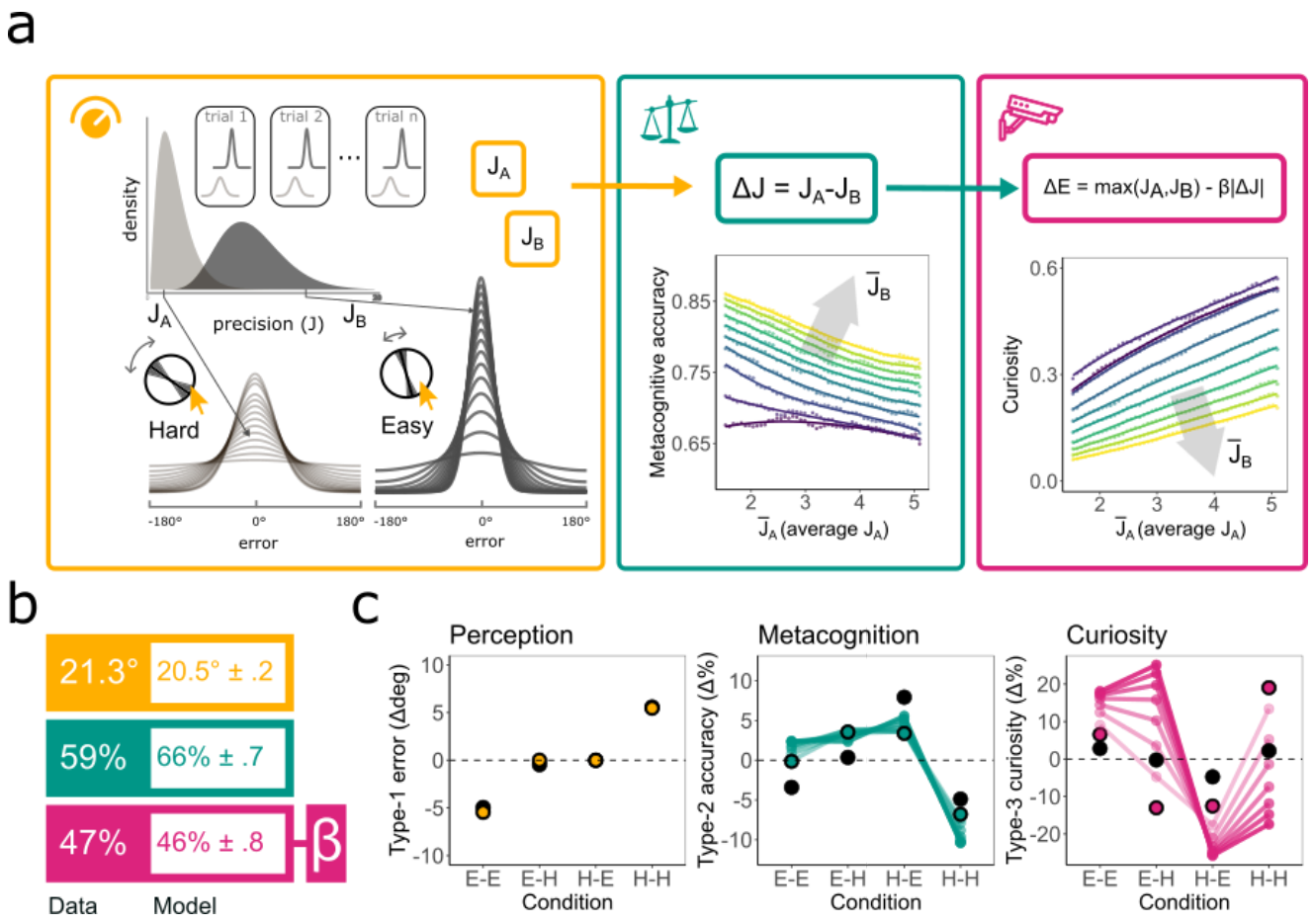

Figure 3. Observer model. **(a) Left 'perception' panel (yellow):** For each trial, the response error is drawn from a circular normal distribution (Von Mises), centred on the correct orientation as shown in the bottom figures. The internal precision of the representation varies across trials according to a Gamma distribution. The top figure illustrates how the Gamma distribution changes with the scale parameter. Light and dark grey represent hard and easy trials, respectively. **Middle 'metacognition' panel (blue):** Simulations for an ideal observer. The observer assesses and compares the precision of each trial, preferring the one with greater precision. The blue panel plots the proportion of correct metacognitive judgments (y-axis) against the difference in average perceptual precision for Trial A (x-axis). The change in precision is achieved through the scale of the Gamma distribution. The colour gradient shows how metacognitive accuracy shifts with an increase in Trial B's average precision relative to Trial A following a change in scale. Note that the 'proportion correct' indicates the highest level of performance attainable given the limits of perceptual sensitivity. **Right 'Curiosity' panel (pink):** Simulations for an ideal observer. Here, the observer decides to focus on metacognition (or otherwise on perception) by weighing the perceptual evidence of the chosen trial ($\max(J_A, J_B)$) against the metacognitive evidence (ΔJ). To arbitrate between these, metacognitive evidence must be scaled by a fixed proportionality factor ($\beta = 1.42$, see main text). This diagram displays the average curiosity for metacognition as a function of average perceptual precision in each trial of a pair (J_A, J_B). **(b)** This section details the sensitivity for perception, metacognition, and curiosity, alongside the values derived from the model, β represents the curiosity scaling factor. **(c)** Displays the relative sensitivity for each decision order as forecasted by the observer model. Black-circled coloured dots indicate ideal observer model predictions (with 500-iteration bootstrapped SEM, error bars are all smaller than the dots), based exclusively on perceptual responses from experiments. Black dots show the empirical group averages from experiments.Coloured dots and lines represent predictions for an observer model with different metacognitive bias (darker colour for greater bias in the [0,2] range).

parameter of the Gamma, metacognitive accuracy steadily decreases with increasing average precision, and curiosity for metacognition increases. Our sample size made the fit per participant impractical, we therefore used an aggregated observer model approach, while keeping assumptions about noise structure to the minimum to avoid grouping artefacts.

The average precision range was selected considering the observed precision values in our empirical data. **Figure 3b-c** shows the observed error, metacognitive accuracy and curiosity rates, and the predicted rate from fitting the model to perceptual responses in the Hard and Easy trials separately. The observer is deemed ideal for metacognition since no additional noise is added to the evidence signal. The curiosity decision is more subjective: it involves comparing metacognitive and perceptual evidence. By design, metacognitive evidence will always be lower than the perceptual evidence, the latter being calculated and the former's difference. A scaling factor is required, which reflects the curiosity for metacognitive feedback. Here, we consider the probability of selecting metacognitive feedback at the group level, which was not significantly different from 50%. We identified the most robust scaling factor across a variety of realistic perceptual evidence distribution ($\beta = 1.42$, see Supplementary Material). Our model has 4 parameters (the scale and shape of the Gamma for Hard and Easy trials), and a fixed scaling factor for curiosity. Our model qualitatively predicts perceptual accuracy across and between conditions, and also captures certain trends in metacognitive accuracy (**Figure 3b**). Yet, metacognitive accuracy was higher for the ideal observer model, suggesting suboptimal metacognitive decisions, a recurring finding in the literature (Mamassian, 2016). An aspect that is not predicted by the model is the difference between H-E and E-H conditions: while the ideal observer predicts no difference, the empirical data departs from this, showing a marked increase in metacognitive accuracy for the H-E condition, probably also mirrored by curiosity. As shown in **Figure 2c-e**, this is attributable to a systematic metacognitive recency bias. Regarding curiosity, the model does not capture the E-H vs. H-E imbalance either. A biased observer model allows for such variability in both metacognition and curiosity (Figure 3c, coloured lines, see Supplementary Material for more details).

Discussion

In his work *On Being a Busybody (De curiositate)*, Plutarch extols the benefits of shifting one's curiosity from things outwards and "turning it inwards" (Plutarch, 1939 ed., p. 477): only those who actively probe their cognition shall ultimately understand others. During planning, acting, or while exercising restraint, human cognition must systematically adjust to multiple sources of external and internal uncertainty. In a similar vein to how external uncertainty can be reduced by enhancing the predictive power of the world model, internal uncertainty may — to some reasonable extent — be mitigated by improving the model an agent has of themselves. In the current work, we investigated the comparative curiosity for feedback on the accuracy of perception and metacognition. Confidence sways our everyday decisions, fuelling our prudence and motivation to change. Yet, it is usually feedback about the outcome of a decision, rather than about one's confidence while making the decision, that seems to be the primary target of self-evaluation. We designed a new paradigm that allowed us to

dissociate the interest for each of these dimensions, building on the realistic assumption that confidence is often used to arbitrate between subsequent decisions and direct our actions.

Our main finding is the existence of an inherent curiosity drive about metacognitive ability, which, under adequate circumstances, prevails over a curiosity about perceptual ability and decision outcome. The finding that individuals are ready to sacrifice knowledge of their gain in a specific trial to confirm the accuracy of their metacognitive skills offers compelling support for the idea that humans greatly value metacognitive knowledge. We designed our study to present participants with a choice: to know the outcome of a specific decision they made or to assess the accuracy of their metacognition without knowing the actual result. This approach intentionally separated cognitive and metacognitive information to explore their interactions. A practical, real-life illustration is seeking advice after self-medicating. A patient who selects one medication over another does so without confirmation of that choice being the safest or most effective. Even if they experience relief, doubts about the decision's correctness and safety persist. Crucially, they may question whether their confidence in the decision was justified. Seeking medical advice in that context would therefore yield metacognitive information beyond the primary decision outcome, significantly influencing future decision-making.

One important aspect of our findings is the notable interindividual variability in this curiosity about metacognition (**Figure 2a**). Previous research has uncovered a variety of motives for human curiosity that transcend the direct value of resolving uncertainty (Kobayashi et al., 2019). Furthermore, overall intrapersonal curiosity has been found to vary across the population, with greater curiosity often associated with a heightened sensitivity to other people's expressions, increased levels of distress, and a concern regarding the most effective ways to cope with worry (Litman, Robinson & Demetre, 2017). Metacognitive representations themselves have been described as at least partially resulting from cultural adaptation, suggesting a rich, varied phenotype of metacognitive traits across individuals and communities (Heyes et al., 2020). Therefore, identifying the specific personality features, cultural underpinnings and neurotypical factors influencing curiosity about metacognition could be crucial in enhancing individualised interventions.

Beyond individual differences, we identified a correlation between curiosity for metacognition and both the difficulty and quality of metacognitive judgement. This link was notable in incentivised experiments (Exp 1 and 3), hinting at the impact of payoff structures on metacognitive monitoring (Locke et al., 2020). By orthogonally manipulating perceptual and metacognitive evidence, we could discern the influence of each evidence type on curiosity. Participants displayed increased curiosity about their perception in trials with high metacognitive evidence ('E-H' and 'H-E' conditions), keen to know their perceptual decision outcome. Conversely, in scenarios with lower metacognitive evidence ('H-H' and 'E-E'), curiosity towards metacognition rose. A similar pattern emerged for metacognitive accuracy across different conditions: incorrect metacognitive decisions led participants to prefer feedback on metacognition rather than perception, suggesting an active monitoring. Implementing such a monitoring and error detection at the metacognitive level could be facilitated via a reliability estimate of metacognitive computations (Recht et al., 2022). Our observer model (**Figure 3**) illustrates how sensory evidence may be gathered, evaluated, and utilised in decisions related to metacognition and curiosity, while also highlighting how metacognitive bias may affect the pursuit of feedback.

Some metacognitive inefficiencies are systematic, while others stem from unpredictable noise-driven variability (Shekhar & Rahnev, 2021). Attention, for example, is known to systematically affect metacognition (e.g., Recht et al. 2019, 2023). The literature distinguishes between metacognitive bias and sensitivity: bias indicates average behaviour shifts across evidence levels, whereas sensitivity measures the metacognitive signal's responsiveness to changes in primary evidence (Fleming & Lau, 2012). Metacognitive bias, being systematic, is tied to specific environments, conditions, or agents and remains relatively stable within them. Both bias and sensitivity influence metacognitive evaluations, with our accuracy measure integrating these aspects into a unified metric, both ultimately having real-life consequences. However, our experimental design also allowed us to uncover a metacognitive bias towards the second decision in a pair (**Figure 2 c-e**). Intriguingly, curiosity appeared to respond to this bias qualitatively: with higher curiosity when the environment contradicted the bias and lower when it reinforced it (Figure 2c). While further replication is needed, this suggests that observers can detect subtle imbalances in metacognitive evidence over time, choosing to monitor metacognitive performance more closely in unexpected environments. Our findings propose that variations in metacognitive abilities may stimulate curiosity and monitoring, paving the way for comprehensive research into metacognitive self-supervision and enhancement.

Inefficient decision-making is often seen as correctable through effective monitoring and control. Confidence, for example, has been suggested to act as an internal learning signal in the absence of external feedback (Ptasczynski et al., 2022; Guggenmos et al., 2016). However, if supervisory processes themselves are inefficient, they might contaminate the overall decision process and heighten the risk of failure. Consequently, metacognitive processes too may well benefit from monitoring and fine-tuning in certain circumstances. This form of meta-control can be facilitated by reallocating cognitive resources 'on-the-fly' among different decision orders, such as from metacognition to meta-metacognition (Recht et al., 2022), thereby extending the interplay between cognition and metacognition (Rosenbaum et al., 2022; Maniscalco et al., 2017). This 'nested cognition' aligns with recent theories suggesting the brain may employ meta-learning algorithms, notably within the prefrontal cortex (Wang et al., 2018). The acceleration in learning empirically observed with increased experience implies that the learning process itself could be optimised across different timescales, a form of ‘learning to learn’ (Wang, 2021). A plausible explanation for this phenomenon is the acquisition of inductive biases—constants that enhance the transferability of learning strategies and categorisation to novel concepts. Another potential aspect of meta-learning could be the refinement of the monitoring and control systems. The discovery of an adaptive curiosity towards metacognitive abilities provides strong evidence for a motivation to enhance the supervisory elements of reasoning, the success in deploying efficient cognitive algorithms being contingent upon the quality of supervision, control, and adaptation of predictive models (Gershman et al., 2015).

Finally, curiosity has been portrayed as an intrinsic motivational force: an appetite for information without an extrinsically rewarding goal (Gottlieb & Oudeyer, 2018; Berlyne, 1954). Other theorists argued for a more liberal definition, considering both extrinsic and intrinsic aspects of curious exploration as irrevocably linked (Kidd & Hayden, 2015). The curiosity drive regarding metacognition found in the present study may well reflect a plurality

of distinct motives, encompassing both an urge to know oneself and a potential desire to improve one's decisions in the long run. How these features evolve and interact across contexts, curriculums, and agents is a relevant question for further studies. Overall, our work contributes to a better understanding of self-regulation by highlighting a systematic and adaptive curiosity about one's own metacognitive abilities and suggesting a way to study its function.

Acknowledgements

This work was supported by a Fyssen Foundation fellowship and a grant from Pembroke College to SR, and a Yinghua Scholarship from Tsinghua University to YY. The authors would like to thank Monica Barbir, Matan Mazor and Jacques Pesnot Lerousseau for helpful comments on this project.

Authors contributions

Conceptualisation: SR, CL & YY; Data curation: SR, CL & YY; Formal Analysis: SR & YY; Writing—original draft: SR, CL, YY & KC; Writing—review & editing: SR, CL, YY & KC. Supervision & funding acquisition: SR.

References

Aguilar-Lleyda, D., & de Gardelle, V. (2021). Confidence guides priority between forthcoming tasks. *Scientific Reports*, *11*(1), 18320. https://doi.org/10.1038/s41598-021-97884-2

Ahmed, A. S., & Duellman, S. (2013). Managerial overconfidence and accounting conservatism. Journal of accounting research, 51(1), 1-30.

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012). What failure in collective decision-making tells us about metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1350-1365. https://doi.org/10.1098/rstb.2011.0420

Berlyne, D. E. (1954). A theory of human curiosity, *British Journal of Psychology*, 45:3

Boldt, A., Blundell, C., & De Martino, B. (2019). Confidence modulates exploration and exploitation in value-based learning. *Neuroscience of Consciousness*, *2019*(1). https://doi.org/10.1093/nc/niz004

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn* (Vol. 11). Washington, DC: National academy press.

Buratti, S., & Allwood, C. M. (2015). Regulating metacognitive processes—support for a meta-metacognitive ability. Metacognition: Fundaments, Applications, and Trends: A profile of the current State-of-the-Art, 17-38.

Cox, M. T. (2005). Metacognition in computation: A selected research review. *Artificial Intelligence*, *169*(2), 104–141. https://doi.org/10.1016/j.artint.2005.10.009

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological science*, *29*(5), 761-778.

Fisher, P. L., & Wells, A. (2008). Metacognitive therapy for obsessive–compulsive disorder: A case series. *Journal of behavior therapy and experimental psychiatry*, *39*(2), 117-132.

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1280-1286.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in human neuroscience*, *8*, 443.

Fu, T. S. T., Koutstaal, W., Poon, L., & Cleare, A. J. (2012). Confidence judgment in depression and dysphoria: The depressive realism vs. negativity hypotheses. *Journal of behavior therapy and experimental psychiatry*, *43*(2), 699-704. https://doi.org/10.1016/j.jbtep.2011.09.014

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273-278.

Geurts, L. S., Cooke, J. R., van Bergen, R. S., & Jehee, J. F. (2022). Subjective confidence reflects representation of Bayesian probability in cortex. *Nature Human Behaviour*, *6*(2), 294-305.

Gottlieb, J., & Oudeyer, P. Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, *19*(12), 758-770.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357-364.

Grinblatt, M., & Keloharju, M. (2009). Sensation seeking, overconfidence, and trading activity. *The Journal of Finance*, *64*(2), 549–578. https://doi.org/10.1111/j.1540-6261.2009.01443.x

Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *Elife*, *5*, e13388.

Händel, M., & Fritzsche, E. S. (2016). Unskilled but subjectively aware: Metacognitive monitoring ability and respective awareness in low-performing students. *Memory & Cognition*, *44*, 229-241.

Hattie, J. (2008). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. Routledge.

Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing ourselves together: The cultural origins of metacognition. *Trends in cognitive sciences*, *24*(5), 349-362.

Iancu, I., Bodner, E., & Ben-Zion, I. Z. (2015). Self esteem, dependency, self-efficacy and self-criticism in social anxiety disorder. *Comprehensive psychiatry*, *58*, 165-171. https://doi.org/10.1016/j.comppsych.2014.11.018

Kaustia, M., & Perttula, M. (2012). Overconfidence and debiasing in the financial industry. *Review of Behavioral Finance*, *4*(1), 46-62. https://doi.org/10.1108/19405971211261100

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449-460.

Kobayashi, K., Ravaioli, S., Baranès, A., Woodford, M., & Gottlieb, J. (2019). Diverse motives for human curiosity. *Nature human behaviour*, *3*(6), 587-595.

Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, *623*(7985), 115-121.

Litman, J. A., Robinson, O. C., & Demetre, J. D. (2017). Intrapersonal curiosity: Inquisitiveness about the inner self. *Self and Identity*, *16*(2), 231-250.

Locke, S. M., Gaffin-Cahn, E., Hosseinizaveh, N., Mamassian, P., & Landy, M. S. (2020). Priors and payoffs in confidence judgments. *Attention, Perception, & Psychophysics*, *82*, 3158-3175.

Lysaker, P. H., Gagen, E., Moritz, S., & Schweitzer, R. D. (2018). Metacognitive approaches to the treatment of psychosis: a comparison of four approaches. *Psychology Research and Behavior Management*, 341-351.

Mamassian, P. (2016). Visual confidence. *Annual Review of Vision Science*, *2*, 459-481.

Mamassian, P. (2020). Confidence forced-choice and other metaperceptual tasks. *Perception*, *49*(6), 616-635.

Maniscalco, B., McCurdy, L. Y., Odegaard, B., & Lau, H. (2017). Limited cognitive resources explain a trade-off between perceptual and metacognitive vigilance. *Journal of neuroscience*, *37*(5), 1213-1224.

Pescetelli, N., Hauperich, A. K., & Yeung, N. (2021). Confidence, advice seeking and changes of mind in decision making. *Cognition*, *215*, 104810.

Plutarch, On Being a Busybody, Vol. VI, Loeb Classical Library edition, 1939.

Ptasczynski, L. E., Steinecker, I., Sterzer, P., & Guggenmos, M. (2022). The value of confidence: Confidence prediction errors drive value-based learning in the absence of external feedback. *PLOS Computational Biology*, *18*(10), e1010580.

Recht, S., de Gardelle, V., & Mamassian, P. (2021). Metacognitive blindness in temporal selection during the deployment of spatial attention. *Cognition*, *216*, 104864. https://doi.org/10.1016/j.cognition.2021.104864

Recht, S., Jovanovic, L., Mamassian, P., & Balsdon, T. (2022). Confidence at the limits of human nested cognition. *Neuroscience of consciousness*, *2022*(1), niac014. https://doi.org/10.1093/nc/niac014

Recht, S., Mamassian, P., & de Gardelle, V. (2019). Temporal attention causes systematic biases in visual confidence. *Scientific Reports*, *9*(1), 11622. https://doi.org/10.1038/s41598-019-48063-x

Recht, S., Mamassian, P., & de Gardelle, V. (2023). Metacognition tracks sensitivity following involuntary shifts of visual attention. *Psychonomic Bulletin & Review*, *30*(3), 1136-1147. https://doi.org/10.3758/s13423-022-02212-y

Rollwage, M., Dolan, R. J., & Fleming, S. M. (2018). Metacognitive failure as a feature of those holding radical beliefs. *Current Biology*, *28*(24), 4014-4021.

Rosenbaum, D., Glickman, M., Fleming, S. M., & Usher, M. (2022). The cognition/metacognition trade-off. *Psychological Science*, *33*(4), 613-628.

Salovich, N. A., & Rapp, D. N. (2021). Misinformed and unaware? Metacognition and the influence of inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(4), 608.

Sawaya, Y., Sharif, M., Christin, N., Kubota, A., Nakarai, A., & Yamada, A. (2017, May). Self-confidence trumps knowledge: A cross-cultural study of security behavior. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 2202-2214).

Schneegans, S., Taylor, R., & Bays, P. M. (2020). Stochastic sampling provides a unifying account of visual working memory limits. Proceedings of the National Academy of Sciences, 117(34), 20959-20968.

Shekhar, M., & Rahnev, D. (2021). Sources of metacognitive inefficiency. *Trends in Cognitive Sciences*, *25*(1), 12-23.

Sherman, M. T., & Seth, A. (2023). Knowing that you know that you know: Above chance discrimination of metacognitive performance. https://doi.org/10.31234/osf.io/qxvf5 (PREPRINT)

Stotz, O., & von Nitzsch, R. (2005). The perception of control and the level of overconfidence: Evidence from analyst earnings estimates and price targets. *Journal of Behavioral Finance*, *6*(3), 121–128. https://doi.org/10.1207/s15427579jpfm0603_2

Van Camp, L., Sabbe, B. G. C., & Oldenburg, J. F. E. (2019). Metacognitive functioning in bipolar disorder versus controls and its correlations with neurocognitive functioning in a cross-sectional design. *Comprehensive psychiatry*, *92*, 7-12. https://doi.org/10.1016/j.comppsych.2019.06.001

Van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological review*, *121*(1), 124.

Van den Berg, R., Shin, H., Chou, W. C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780-8785.

Wang, J. X. (2021). Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, *38*, 90–95. https://doi.org/10.1016/j.cobeha.2021.01.002

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., ... & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, *21*(6), 860-868.

Wells, A., Capobianco, L., Matthews, G., & Nordahl, H. M. (2020). Editorial: Metacognitive therapy: Science and practice of a paradigm. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.576210

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310-1321.

Zheng, Y., Recht, S., & Rahnev, D. (2023). Common computations for metacognition and meta-metacognition. *Neuroscience of Consciousness*, *2023*(1), niad023. https://doi.org/10.1093/nc/niad023