Confidence in metacognition
Type 3 judgments in the context of perceptual decisions

Q. Cavalan, J.C. Vergnaud & V. de Gardelle

September 30, 2024

Abstract

The ability to form judgments about one's own cognitive abilities, known as metacognition, has been extensively studied recently. Metacognitive judgments are not always accurate, but the extent to which individuals can evaluate the quality of their own metacognition remains unclear. To explore this ability, termed meta-metacognition, we present here two experiments. Observers were asked to judge which of two perceptual decisions was more likely to be correct (a metacognitive choice) and to evaluate their confidence in such choices (a meta-metacognitive judgment). By collecting confidence ratings after each perceptual decision and evaluating how individuals make probabilistic inference in a separate task, we demonstrated two key empirical findings. First, meta-metacognitive judgments predicted the quality of metacognition beyond the information that can be optimally extracted from perceptual confidence. Second, participants' metacognitive choices contradicted their ratings of perceptual confidence in nearly 25% of cases, and when this occurred, their metacognitive accuracy mostly improved.

Research Transparency Statement

General Disclosures

Conflicts of interest: All authors declare no conflicts of interest. Funding: This research was supported by Agence Nationale de la Recherche (ANR-18-CE28-0015). Artificial intelligence: No artificial intelligence assisted technologies were used in this research or the creation of this article. Ethics: This research received approval from a local ethics board (IRB 2023-041).

Experiment 1 and 2

Preregistration: The design, exclusion criteria and sample size for both experiments were preregistered (https://aspredicted.org/vt2mq.pdf) on 2022-18-03, prior to data collection. The analysis plan was not preregistered. Materials: All study materials, primary data and analysis scripts are publicly available (https://osf.io/28znj/).

1 Introduction

Metacognition is the ability to monitor and regulate one's own cognitive processes (Nelson and Narens, 1994). One simple form of metacognition is the expression of confidence in our judgments

or the detection of our own errors (Boldt and Yeung, 2015; Yeung and Summerfield, 2012). Such metacognitive evaluations can be instrumental for behavioral adaptation. For instance, realizing that one has chosen a bad project allows stopping it before wasting resources on this project. Detecting an error we have just made can lead us to adopt a more cautious decision strategy on subsequent occasions (Rabbitt, 1966; Gehring et al., 1993). Error signals also play an essential role in learning processes (Holroyd and Coles, 2002), and our inner sense of whether we might have made a mistake can thus guide learning even when no external feedback is available (Guggenmos et al., 2016; Daniel and Pollmann, 2012; Hainguerlot et al., 2018).

However, individuals are not always good at detecting their own mistakes. For instance, recent studies have shown that metacognition is degraded when working memory resources are engaged in complex multi-tasking (Konishi et al., 2021; Maniscalco and Lau, 2015) or when endogenous attention must be deployed (Recht et al., 2021, 2019). Such metacognitive blind spots suggest that metacognition is not completely cost-free but relies on executive resources, whose availability may vary within an individual depending on the context or on external demands.

Here, we reason that if metacognition regulates behavior, and if metacognitive abilities vary within an individual, one key question that arises is whether individuals know when to trust their own metacognition. This question brings us to the notion of meta-metacognition, that is the knowledge individuals have about their own metacognition. Monitoring metacognition can be important: the more certain an individual is of having made a mistake in their work, the more time they are willing to spend in order to identify and correct it.

Yet, only a few studies have investigated the issue of meta-metacognition so far. In the memory domain, Dunlosky et al. (2005) showed that participants could not only evaluate whether they have correctly learnt an item (a metacognitive judgment), but also predict the accuracy of such evaluations (a meta-metacognitive judgment). Buratti et al. (2013) showed that when given the opportunity to correct some of their previous confidence judgments, participants were able to identify the most biased judgments, which the authors interpreted as evidence for meta-metacognition. In the domain of perceptual decisions, three recent studies showed that participants exhibit above-chance sensitivity at the meta-metacognitive level, when evaluating the degree of certainty in a just-given confidence judgment (Zheng et al., 2023) or when indicating which of two previous confidence ratings (made in two distinct trials) better reflected their actual performance (Recht et al., 2022; Sherman and Seth, 2024). However, it has also been suggested that meta-metacognitive judgments could be reducible to metacognitive judgments (Zheng et al., 2023).

The present work offers both methodological and empirical contributions to this emergent literature. At the methodological level, we present an original experimental paradigm where participants report their confidence in their error detection judgments, which allows for an unambiguous definition of metacognitive accuracy on a trial-by-trial basis, enabling a clear formalisation and an appropriate incentivization for both metacognitive and meta-metacognitive judgments. From this, we can derive evidence for meta-metacognition in a model-free approach, directly from observable responses, without any need to estimate unobserved model parameters, and we can define and measure precisely whether participants are underconfident or overconfident regarding their metacognitive ability, i.e. a meta-metacognitive bias, which has not been examined in prior research. At the empirical level, we demonstrate that meta-metacognitive judgments cannot be reduced to metacognitive judgments, showing that meta-metacognition must be considered as a distinct level of processing. Finally, we

document changes of mind at the metacognitive level, where confidence ratings can be appropriately overruled by error detection judgments.

In our experimental studies, participants performed a perceptual task (referred to as the Type 1 task) and whenever two consecutive trials contained one correct and one incorrect answers, they were prompted to indicate which of the previous two trials was the correct one (a metacognitive task, referred to as the Type 2 forced choice task). This procedure allows us to identify unambiguously correct and incorrect metacognitive judgments. Then, participants reported their confidence in this decision, constituting a Type 3 or meta-metacognitive judgment. In Experiment 1, we show that participants are able to form above chance judgments on the quality of their metacognitive monitoring, thereby providing evidence for meta-metacognitive abilities and replicating prior research on the topic (Recht et al., 2022; Sherman and Seth, 2024; Zheng et al., 2023). In Experiment 2, within this paradigm, we additionally collected confidence ratings following each perceptual trial, as well as measures of probabilistic inference based on a separate urn task. By doing so, we could quantify the amount of information regarding the accuracy of the Type 2 forced-choice task already present at the metacognitive level.

2 Results

2.1 Perceptual, Metacognitive, and Meta-metacognitive abilities

Participants' perceptual task was to indicate which of four possible colours was dominant in an array of colored squares. When a pair of trials contained one correct and one incorrect decisions, their metacognitive task was to indicate which of the two decisions was correct (see Figure 1). To maximize the number of such Type 2 forced choice decisions, we calibrated the proportion of the dominant color to obtain 50% of perceptual accuracy in a training session prior to the main task. In the main session, perceptual accuracy was slightly higher ($M_{perf} = 56.12\%$, $SD_{perf} = 8.11\%$). We could confirm that participants had metacognitive access to their perceptual performance: accuracy in the Type 2 forced choice task was on average equal to 72.32%, well above the 50% chance level (one-tailed t-test, $t(52) = 23.22$, $p < .001$).

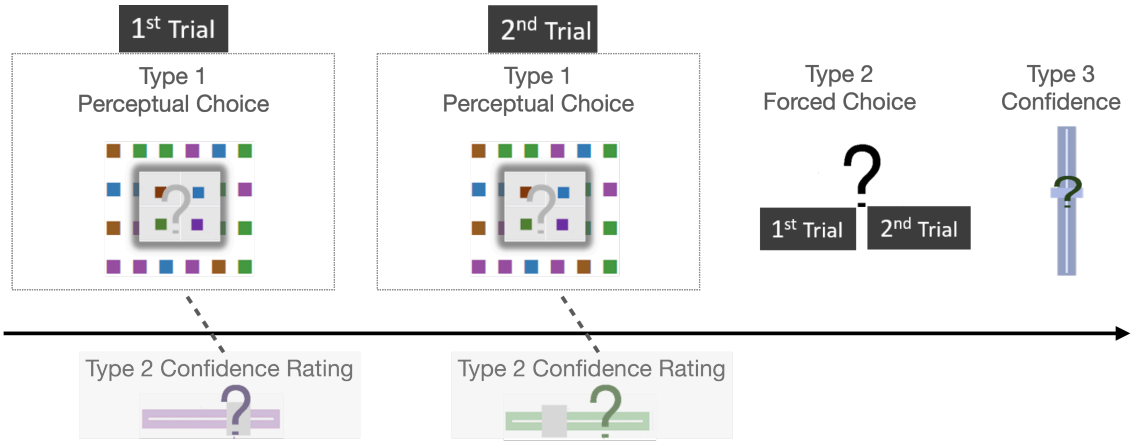

Figure 1: Schematic representation of the experimental design for the main task in Experiments 1 and 2. See Methods for details. Type 2 confidence ratings were only collected for Experiment 2.

To test participants' meta-metacognitive abilities, we evaluated whether Type 3 confidence could discriminate between good and bad Type 2 choices (Figure 2A). Critically, we found that when Type 3 confidence was higher (above the participant's median), then Type 2 accuracy was significantly higher ($M_{high} = 80.30$, $M_{low} = 61.31$, $t(50) = 9.38$, $p < .001$). To go further and avoid relying on a median split, we computed Type 3 sensitivity defined as the area under the Type 3 ROC curve (see Methods), and confirmed that it was higher than chance (Figure 2B, $M_{sensitivity} = 63.77$, $t(52) = 10.14$, $p < .001$). These results demonstrate that participants in our sample exhibited meta-metacognitive abilities: their Type 3 confidence judgments conveyed information about the quality of their Type 2 judgments. Our method also allowed us to evaluate the bias individuals may have when evaluating their metacognition. In this respect, we found that most participants were overconfident at the Type 3 level: their Type 3 confidence was greater than their Type 2 accuracy ($M = 82.09\%$ vs. 72.32%, $t(52) = 6.57$, $p < .001$; Figure 2C), indicating that participants overestimated their ability to identify the correct trial in the pair.

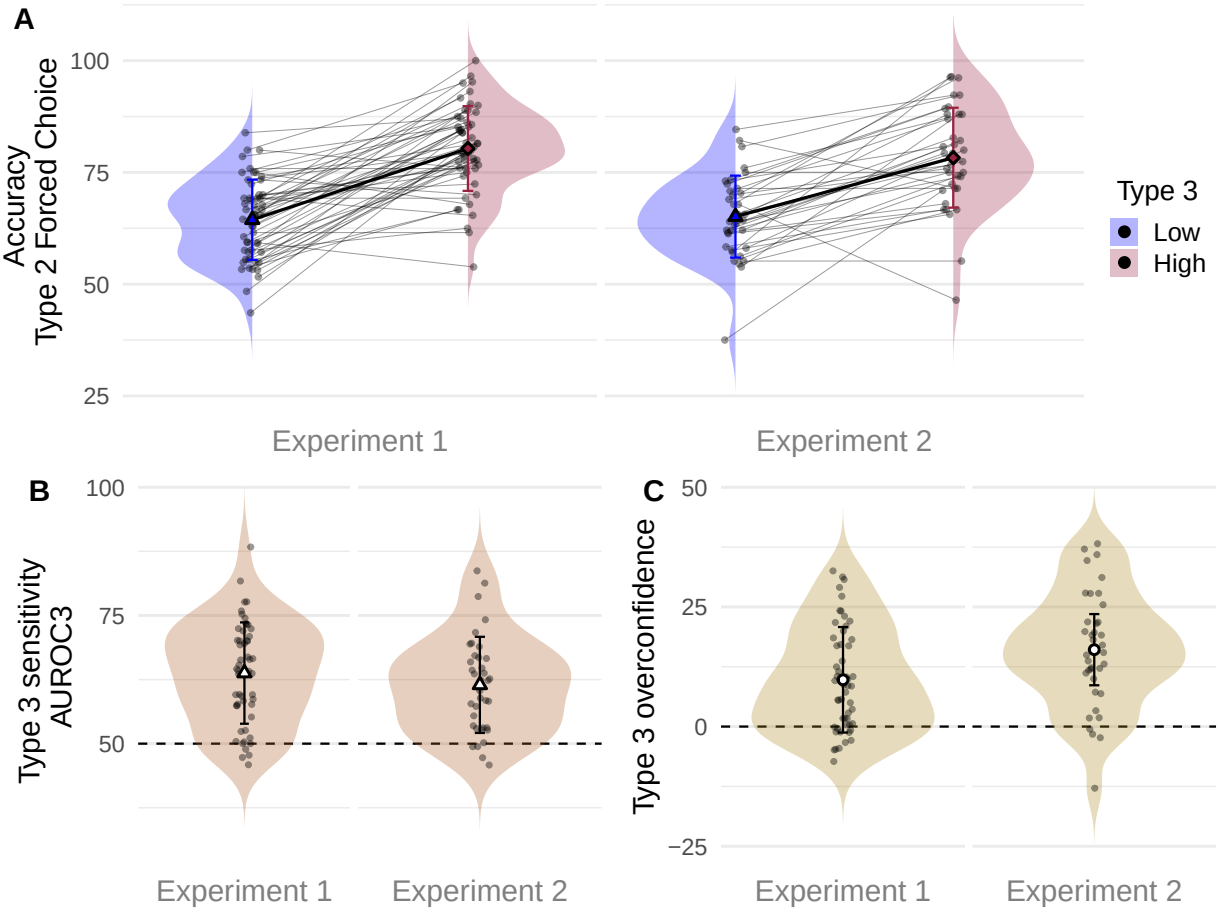

Figure 2: **A**. Distribution of accuracy in the Type 2 forced choice task (proportion of correct Type 2 detection) across participants for low (in blue) and high (in red) Type 3 confidence in Experiment 1 and 2. **B**. Distribution of Type 3 sensitivity (area under the Type 3 ROC curve, see methods for more details) in Experiment 1 and 2. The dotted line corresponds to chance level. **C**. Distribution of Type 3 overconfidence (Type 3 confidence - Type 2 Forced Choice accuracy) in Experiment 1 and 2. The dotted line corresponds to the absence of overconfidence.
Note: Large dots and thick lines represent the mean performance across participants. Error bars represent standard deviation across participants. Small dots and thin lines represent individual data.

2.2 Are Type 3 judgments only inferred from Type 2 ratings?

In Experiment 1, we showed that individuals were able to form informative meta-metacognitive judgments: the accuracy of Type 2 forced choice decisions was indeed predicted from Type 3 ratings. In Experiment 2, our main goal was to replicate this result while controlling for Type 2 confidence ratings after each perceptual decision.

We first replicated the basic findings of Experiment 1 (Figure 2A, 2B and 2C): Type 2 accuracy was significantly higher when Type 3 confidence was high than when it was low ($M_{high} = 78.29$, $M_{low} = 65.12$, $t(36) = 6.87$, $p < .001$), Type 3 sensitivity was greater than chance ($M_{sensitivity} = 61.47\%$, $t(36) = 7.44$, $p < .001$), and Type 3 bias indicated that individuals overestimated their ability to identify the correct trial in the pair (Type 3 confidence: 87.69% vs. Type 2 accuracy: 71.63% ; $t(36) = 8.14$, $p < .001$). Comparing Experiment 1 and 2, we found that the introduction of Type 2 confidence ratings after each trial did not impact perceptual performance, Type 2 accuracy, and Type 3 sensitivity (all p>.2), but led to higher Type 3 overconfidence ($M_{Exp1} = 9.77$ vs. $M_{Exp2} = 16.06$, $t(88) = -2.59$, $p = .011$). In addition, in Experiment 2 we also measured metacognition alone (i.e. without Type 3 evaluations) in an additional set of trials after the main task, and we found that Type 3 judgments did not significantly impact metacognitive performance (Type 2 sensitivity: $M_{with} = 64.44$, $M_{without} = 63.01$, $t(36) = 1.15$, $p = .256$; Type 2 Bias: $M_{with} = 17.31$, $M_{without} = 14.96$, $t(36) = 1.51$, $p = .139$).

Our next analyses aimed at dissecting the computational mechanisms by which participants evaluate their metacognition. As detailed in the Method, Type 2 forced choice accuracy can already be inferred from Type 2 confidence ratings. To evaluate whether participants perform this type of inference, we introduced a new urn task in Experiment 2. This task was designed to be formally equivalent to the Type 3 inference problem. In this urn task, participants had to indicate their belief about whether a green ball came from urn A or urn B. Critically, the probabilities of drawing a green ball for the two urns were set to match the Type 2 confidences measured in a given trial pair in the main task (Figure 3). This allowed us to define, for each participant and in each pair of trials, a pseudo-Type 3 rating (see Methods), which corresponds to the participants' subjective probability that they selected the correct trial in the pair, given their Type 2 confidence ratings. This pseudo-Type 3 is directly comparable to participants' actual Type 3 rating.

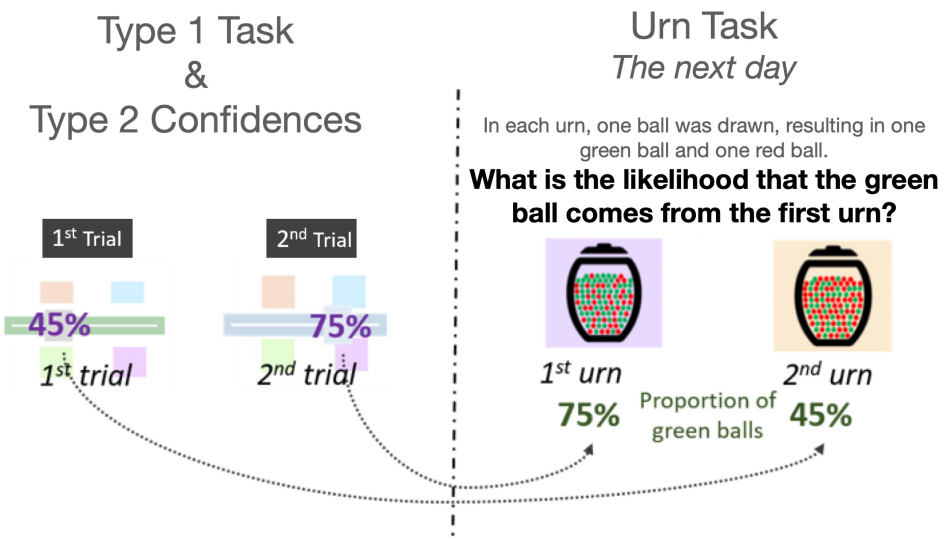

Figure 3: A practical example of the urn task. In a specific pair of perceptual decisions, a participant provided a Type 2 confidence rating of 45% for the first trial and 75% for the second trial. Consequently, in the urn task on the following day, the first urn will contain 75% green balls, while the second urn will contain 45% green balls.

We verified that participants performed well in the urn task, by comparing their responses with the probability obtained through Bayesian inference (Figure 4A and 4C). We found that the two quantities were highly correlated ($M_r = .71$, $SD = 0.27$, $t(36) = 16.25$, $p < .001$), showed no difference on average ($M_{Diff} = -1.72$, $t(36) = -1.02$,$p = .315$), and the regression coefficient between them was not different from one ($M_\beta = .90$, $t(36) = -1.30$, $p = .203$). Participants' estimated probabilities in the urn task, however, could not explain very well their Type 3 confidences in the main task (Figure 4B and 4D). Specifically, Type 3 confidences were only weakly correlated with pseudo-Type 3 ratings ($Mean_r = .34$, $SD = 0.19$, $t(36) = 10.93$, $p < .001$), the regression slope was lower than 1 ($M_\beta = .20$, $t(36) = -23.30$, $p < .001$), and Type 3 confidences were on average higher than pseudo-Type 3 ratings ($M_{Type3} = 87.40\%$ vs. $M_{pseudo-Type3} = 66.81\%$, $t(36) = 11.36$, $p < .001$). This suggests that Type 3 confidence in the main task is partly, but not fully explained by participants applying probabilistic reasoning on their own Type 2 confidence ratings.

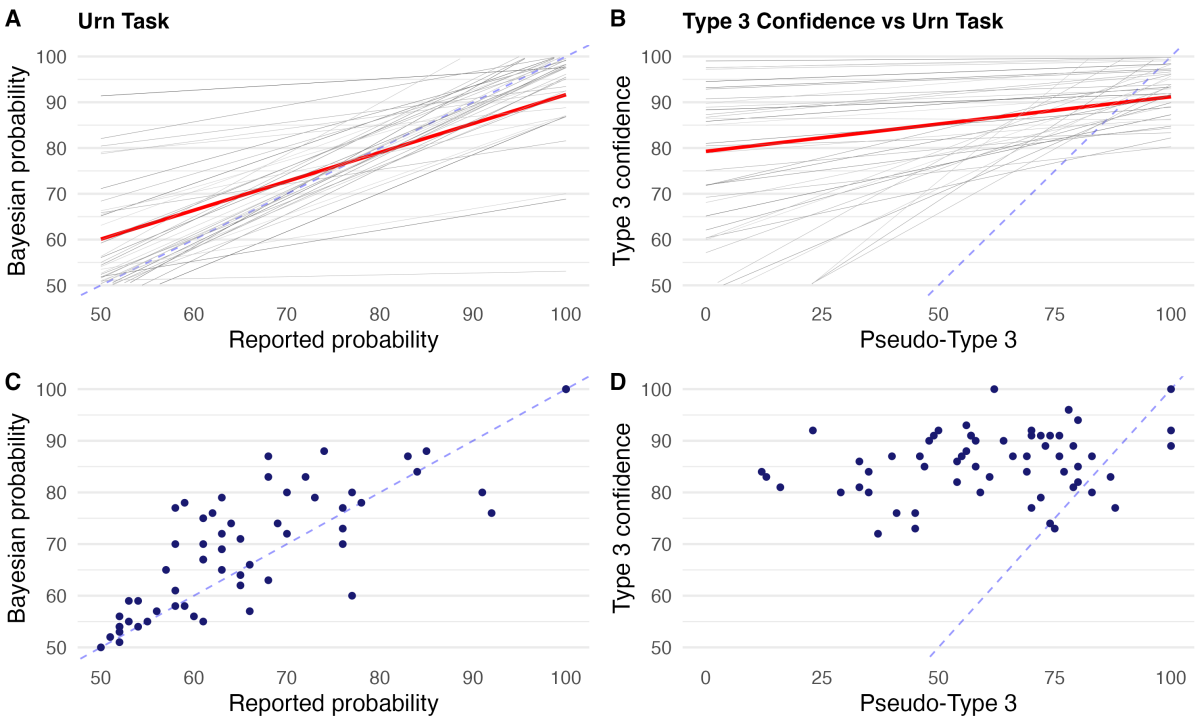

Figure 4: **A**. Bayesian probability plotted against the reported probability in urn task. Gray lines correspond to linear regressions for each participant. The red line corresponds to the linear regression pooling all observations. **B**. Type 3 confidence plotted against pseudo-Type 3 (see methods for more details). **C**. Same as A, but focusing on a single participant. Blue dots represent individual observations. **D**. Same as B, but focusing on the same participant as in C. Blue dots represent individual observations.

So far, we have shown that participants' Type 3 confidence judgments predict their accuracy in the Type 2 forced choice, and are only weakly explained by pseudo-Type 3 ratings. Our next goal was then to evaluate whether Type 3 judgments could constitute an independent source of information regarding Type 2 accuracy, above and beyond the information that is contained in the pseudo-Type 3 ratings. To do so, we compared regression models in which the accuracy of the participants' Type 2 forced choices was predicted from both their Type 3 confidence and pseudo-Type 3 ratings, or from only one of these variables. The best-fitting model ($AIC = 2254$) included a fixed intercept, a fixed main effect for each parameter and a random intercept at the participant level (see Supplementary Materials for the regression table). Importantly, in this model both Type 3 confidence ($z = 5.84$, $p < .001$) and pseudo-Type 3 ($z = 5.83$, $p < .001$) were significant predictors of the correct Type 2 choice. In other words, participants' Type 3 confidence was predictive of whether their Type 2 detection was correct or not, even when we controlled for the information already contained in their Type 2 confidence ratings (as captured by the pseudo-Type 3 variable). Figure 5A illustrates this: not only Type 2 accuracy (y-axis) increases with pseudo-Type 3 ratings (x-axis), but it also increases with Type 3 confidence (red vs. blue dots), when pseudo-Type 3 is kept constant. Note that this analysis also shows that Type 2 ratings contain some information that is no longer present in Type 3 ratings, because pseudo-Type 3 remains a significant predictor in this regression. In the next section, we delve further into this issue, and evaluate how much different predictors can be used to improve Type 2 forced choices.

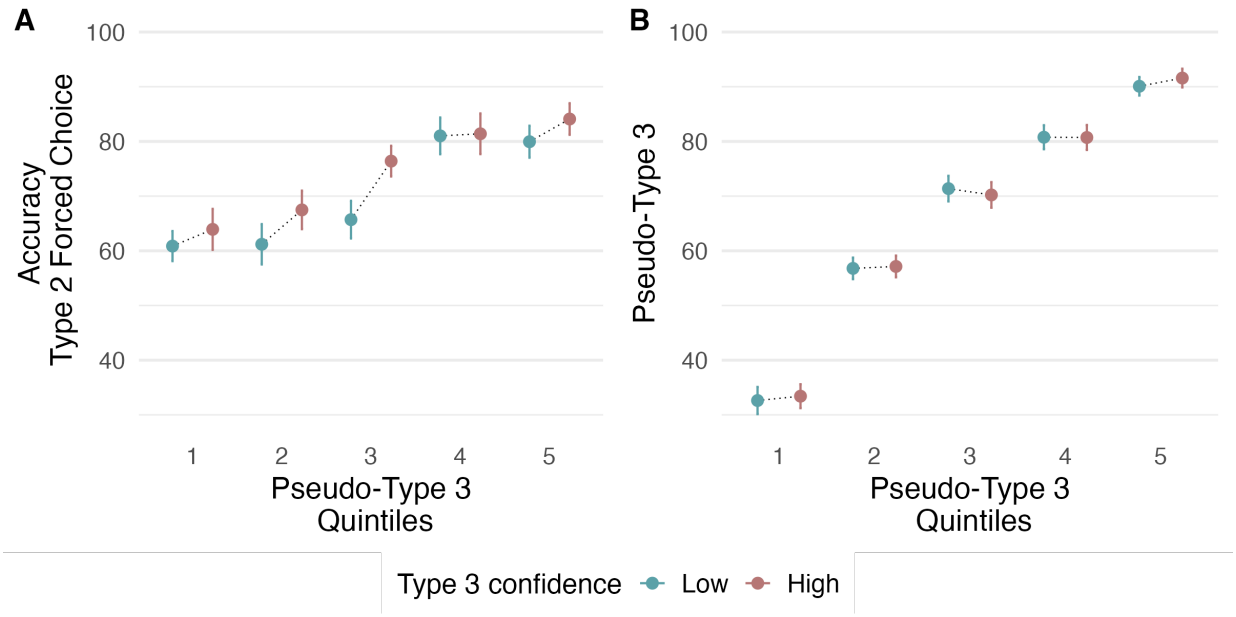

Figure 5: **A**. Accuracy in the Type 2 forced choice task (proportion of correct Type 2 detection) as a function of the pseudo-Type 3 (divided into quintiles) and Type 3 confidence (a median split within each quintile), showing that for any level of pseudo-Type 3, higher Type 3 confidence consistently predicts greater accuracy in the Type 2 forced-choice task. **B**. Pseudo-Type 3 value (in the same subsets as in A) confirm that the observed difference between low and high Type 3 confidence is not due to confounding differences in pseudo-Type 3. In both panels, error bars represent mean and s.e.m. across participants.

2.3 Leveraging information to improve Type 2 choices

Figure 6 illustrates Type 2 forced choice accuracy for participants and for four simulated agents. For these agents, Type 2 choices were based on participants' Type 3 ratings (Type 3 agent), participants' pseudo-Type 3 derived from the urn task (Pseudo 3 agent), selecting the trial with the highest Type 2 rating (Max 2 agent), or they were based on all of these variables (Full agent). For each simulated agent, we searched for the linear combination of predictors (including an intercept) which maximized the accuracy of the agent's Type 2 choice.

All simulated agents performed well above chance. When a single predictor was used, Type 3 information was better at predicting which trial was correct (73.93%) than pseudo-Type 3 (71.18%), which in turn was more predictive than taking the trial with maximum confidence (68.24%). Using all 3 predictors in the full agent greatly increased accuracy (to 78.15%). This full agent also outperformed participants (71.87%), which indicates that in theory one could improve Type 2 choices given the information provided at the different stages by participants. We note also that participants did better than just selecting the trial with the highest Type 2 rating. This last result implies that Type 2 forced choices were not always consistent with Type 2 confidence ratings, a phenomenon which we explore further in our next section.

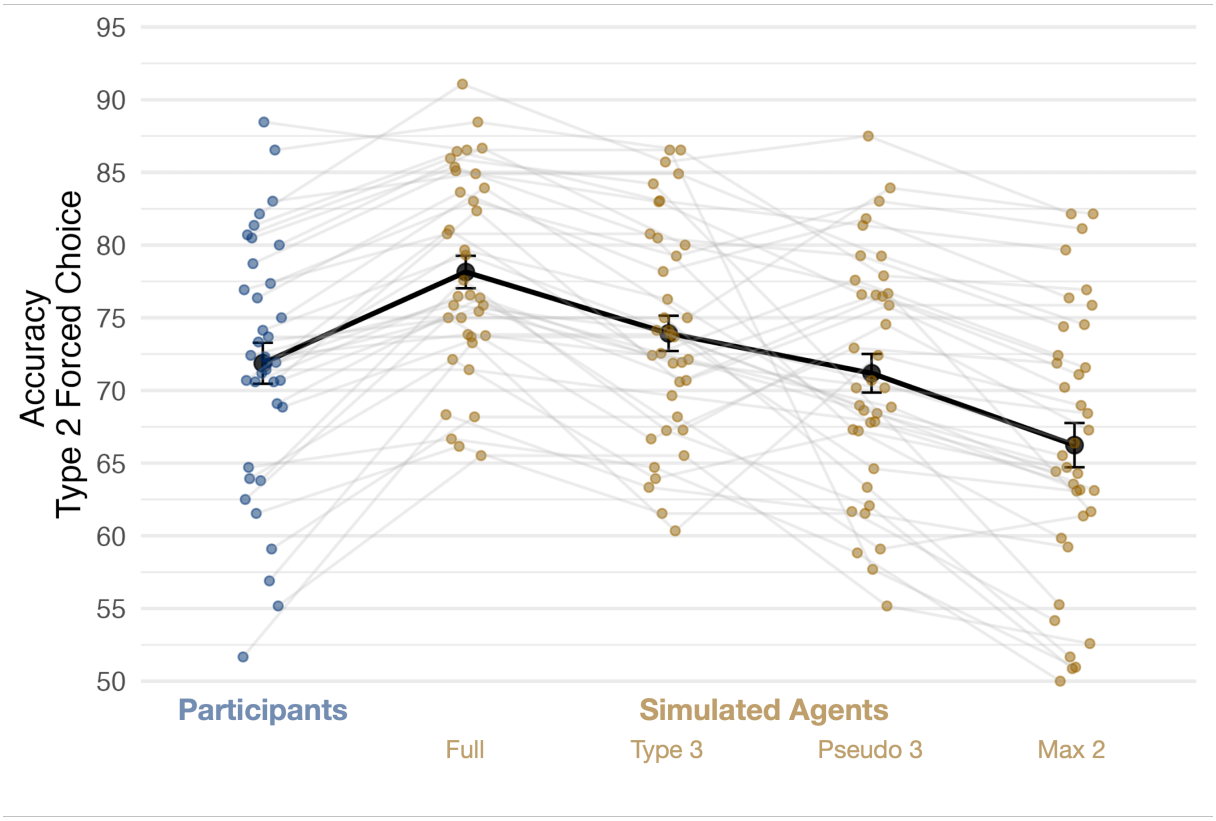

Figure 6: Distribution of accuracy in the Type 2 forced choice task (proportion of correct Type 2 detection) for participants and for simulated agents using various types of information for decision-making: Type 3 ratings (Type 3 agent), pseudo-Type 3 (Pseudo 3 agent), which trial received the highest Type 2 ratings (Max 2 agent) or a combination of these three information sources (Full agent). See methods for more details. Large dots and thick lines represent mean and error bars represent s.e.m, across participants. Small dots and thin lines represent individual data.

2.4 Metacognitive inconsistencies

One would expect participants to consistently select as the correct trial the one for which they were most confident. However, surprisingly, this was not always the case in our data: participants chose the trial for which they were less confident in 23% of the trial pairs. These inconsistent Type 2 decisions occurred mostly when the difference between the two Type 2 confidence ratings was low (the two Type 2 confidence ratings were close) than when it was high, as defined by a median split for each participant ($M_{Low} = 32.09\%$, $M_{High} = 13.80\%$, $t(36) = -9.37$, $p < .001$; Figure 7A). Most importantly, these inconsistencies could not be attributed to a lack of attention and purely random responses, as participants' accuracy for those decisions was above chance ($M_{accuracy} = 57.13$, $Chance = 50$ $t(36) = 2.10$, $p = .044$; When weighted by the number of inconsistent decisions, which varied across participants: $M_{accuracy} = 59.58$, $t(36) = 3.38$, $p = .002$; Figure 7B). Note that this implies that complying with Type 2 ratings would have actually produced a Type 2 accuracy below chance in these inconsistent cases. Thus, those inconsistent decisions were also warranted because they provided a benefit in metacognitive accuracy over simply complying with Type 2 ratings. Across participants, the share of inconsistent decisions was also negatively correlated with

Type 2 confidence sensitivity ($Cor = -.43$, $t(35) = 8.73$, $p = .007$), suggesting that participants may have rightly realized that their Type 2 confidences could be particularly noisy, and if so, may have overridden these ratings when prompted for a Type 2 forced choice.

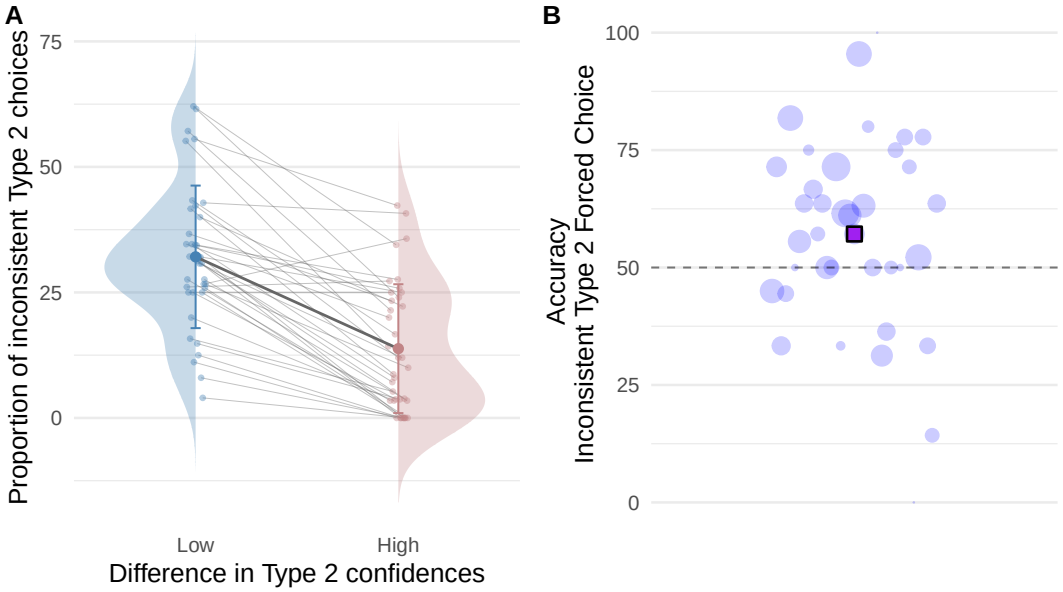

Figure 7: **A**. Distribution of the frequency of inconsistent Type 2 forced choice for low and high difference between the two Type 2 confidence ratings. Large dots and thick lines represent the mean frequency across participants. Error bars represent standard deviation across participants. Small dots and thin lines represent individual data. **B**. Distribution of accuracy in the Type 2 forced choice task (proportion of correct Type 2 detection) for the subset of inconsistent Type 2 forced choices. Dots represent each participant with varying sizes depending on the number of inconsistent decisions they made (between 1 and 29). The square represents the mean accuracy across participants. The dotted line corresponds to chance level.

3 General discussion

In the context of a perceptual task, we explored the characteristics of meta-metacognition, by probing participants' confidence regarding metacognitive judgments (here, a Type 2 forced choice). Critically, in comparison to prior work, our methodology allowed us to define metacognitive accuracy and meta-metacognition unambiguously on each trial, and to evaluate Type 3 sensitivity while controlling for Type 2 information. We observed that meta-metacognitive judgments (Type 3 confidence) carried relevant information, and could not be reduced to Type 2 information. More precisely, Type 3 ratings included additional pieces of information while lacking some elements present in Type 2 confidence ratings. In other words, the different evaluations made by participants were partly complementary. As a result, a simulated observer utilizing both Type 2 and Type 3 information could outperform participants' Type 2 forced-choice accuracy.

On the one hand, some information has been lost at the time of Type 3 ratings: although Type 3 judgments contain information about the accuracy of the Type 2 choice, they did not contain all the information that could have been used. In our regression analysis, for instance, Type 2 ratings

remained a significant predictor of Type 2 choice accuracy, above and beyond Type 3 ratings. To account for the loss of information between the initial evaluations of performance and subsequent ones, one could hypothesize that meta-metacognition involves a readout of the evidence used to form the initial metacognitive evaluation of performance, and that this readout could be corrupted by noise or memory loss, relative to the original evidence.

On the other hand, we also found an information gain from Type 2 ratings to subsequent evaluations. In our data, the inconsistencies between Type 2 confidence ratings and Type 2 forced choices are a striking demonstration of this information gain. These inconsistencies occurred at a relatively high rate (in 25% of the cases in Experiment 2) given that these judgments were made in the same pair of trials, within seconds, but critically they were mostly judicious as they resulted in improved performance in Type 2 accuracy (relative to just taking the decision in the pair associated with the highest confidence). Several explanations can be delineated to explain such changes-of-mind at the metacognitive level. First, one could speculate that participants monitor internal or external factors that affect the quality of their metacognitive evaluations, such as the effort deployed or the trade-off between speed and accuracy used for these evaluations, or the occurrence of a distraction on a given trial. By monitoring such factors, participants could know which metacognitive judgments should be trusted more. Second, it is also conceivable that participants continue to process information about the trial even after the perceptual decision and the Type 2 rating are reported. If they then realize that their perceptual decision on the last trial was incorrect, they would also want to reconsider the confidence rating given on that trial. A third explanation has been put forward, by which metacognitive noise affecting the initial evaluation of confidence could be reduced in a subsequent re-evaluation (Elosegi et al. (2023)).

We note, in passing, that this information gain also echoes certain cases reported in the literature, where Type 2 information (as measured by meta-d') can exceed Type 1 information (as measured by d'). This phenomenon has been frequently observed in the domain of memory (Mazancieux et al., 2020; Ye et al., 2019; Lee et al., 2018) and investigated in the domain of perception as well (Charles et al., 2013; Mamassian and de Gardelle, 2022). Several explanations have been proposed for this observation. A potential explanation is that variations in performance exist, not accounted for in Type 2 analyses but accessible to participants themselves. These variations may stem from internal fluctuations in effort or attention, or external manipulations of task difficulty (e.g., due to staircase procedures, as highlighted in Rahnev and Fleming (2019)). Another explanation suggests that Type 1 responses are suboptimal compared to what participants can achieve because they are made under time pressure, and subsequent Type 2 judgments can reveal more information (Charles et al., 2013). It is also proposed that after their Type 1 decision, participants continuously accumulate evidence at the metacognitive level (Pleskac and Busemeyer, 2010), such that Type 2 judgments made at a later point in time may contain more information. We suggest that such explanations can be also examined at the level of meta-metacognition.

In our methodology, participants report meta-metacognitive judgments only in cases where they have one correct and one incorrect perceptual answer in the series. One could argue that this might raise concerns about the possibility that participants, as they are informed when these cases occur, may learn about their perceptual performance or about their metacognition during the experiment. Since this was more an unintentional consequence of our methodology than a desired choice, we checked whether this feedback on perceptual performance could have somewhat improved partici-

pants' meta-metacognitive abilities. To do so, we ran an additional experiment ($N = 27$) where instead of performing the Type 2 forced choice and Type 3 confidence rating only in some cases, participants performed those tasks after each pair of perceptual decisions, without any feedback regarding perceptual performance. In practice, every two trials, participants were asked to choose the answer that was most likely to be correct and to give their conditional Type 3 confidence in their choice, in the event that they had only one correct answer in the series. The results from this experiment are presented in Supplementary Materials. In a nutshell, we replicate our main findings, and we find that Type 2 and Type 3 sensitivity are virtually unchanged in this supplementary experiment compared to Experiment 2. This suggests that it is unlikely that our method artificially boosted metacognition or induced meta-metacognitive abilities.

To conclude, we will highlight possible avenues for future research on the issue of meta-metacognition. Following the distinction between monitoring and control presented by Nelson and Narens (1994) for metacognition, we (and others) have so far investigated the monitoring aspect of meta-metacognition. This leaves open the question of how such meta-metacognitive judgments might be used by individuals to regulate their behavior. We see at least two distinct possibilities. First, individuals should rely on their own metacognition only when it is deemed reliable. For instance, in the context of collective decision making, individuals should ideally combine their initial choice, weighted by their confidence, to make the best collective response. However, if meta-metacognition indicates that such confidence ratings are at chance, the group could instead resort to a majority rule, disregarding Type 2 estimations. Second, meta-metacognition could be useful to prompt individuals to adjust their behavior so as to improve their metacognitive process. A low Type 3 rating may motivate individuals to exert more cognitive effort in the metacognitive task, or to seek external feedback in place of internal feedback, as mentioned in our introductory example of a student with poor metacognition who would seek external evaluations from a professor. Alternatively, this could be done by changing the context in which individuals operate, e.g. by avoiding factors that contribute to lower metacognition, such as stress (Reyes et al., 2015, 2020) or complex multi-tasking (Maniscalco and Lau, 2015; Konishi et al., 2021). Our study shows that meta-metacognitive reports are able to predict the accuracy of Type 2 choices. Whether individuals are able to harness this meta-metacognitive knowledge to improve their behavior remains an open empirical question.

4 Method

The design, exclusion criteria and sample size for both experiments was pre-registered. The pre-registration can be accessed at https://aspredicted.org/vt2mq.pdf.

Participants

We recruited 58 adults in Experiment 1, and 44 in Experiment 2, using the online database of the Parisian Experimental Economics Laboratory (LEEP). Participants were naïve with respect to the goal of the study, and provided informed consent before the experiment. The research was approved by the institutional review board of Paris School of Economics (application 2023-041).

In Experiment 1, our sample size was determined so that we would be able to test whether

meta-metacognitive sensitivity (as measured by Type 3 ROC, see below) would be above chance level (50%). Since no prior data on Type 3 ROC was available, we relied on prior data from our laboratory where Type 2 ROC was equal to .62 on average ($SD = .069$). We anticipated that Type 3 ROC would be closer to chance and more variable (i.e. $M = .56$, $SD = .092$) than Type 2 ROC (see pre-registration for more details). The anticipated effect size for our main test was thus $d = .535$, and a power analysis revealed that 40 participants would be needed to detect this effect in a one tailed t-test with 95% power and 5% error probability. We increased this number to 60 as we expected a 33% drop-out rate between the first and second session (in practice, the drop-out rate was 3%).

In Experiment 2, a power analysis indicated that 8 participants would be needed to detect the effect size ($d = 1.39$) found in Experiment 1 (one-tailed t test, with 95% power and 5% error probability). Our sample size was well above this number, and aimed at being close to the number of of participants in Experiment 1, for comparison purposes.

Following our pre-registration, we removed data from 12 participants (5 in Experiment 1, and 7 in Experiment 2) before our main analyses, because their performance at the Type 2 task was not significantly higher than the chance level (one sided binomial test against 50%). As a result, our analyses are based on the data of 53 participants for Experiment 1 and 36 participants for Experiment 2.

Design and session

In Experiment 1, our paradigm (Figure 1) involved three types of responses: first, participants completed a perceptual task (Type 1 cognitive task), then they were asked to detect their errors in this perceptual task (Type 2 metacognitive task) and finally to rate their confidence regarding this error detection judgment (Type 3 meta-metacognitive task). In addition, in Experiment 2 participants provided a confidence rating after each perceptual decision.

Experiment 1 was divided into 2 sessions of 30 and 45 minutes respectively, with a first session where participants received training in all the tasks, and a second session where participants completed the main part of the experiment (120 series of 2 perceptual trials). Experiment 2 was made of 3 sessions of 35, 50 and 75 minutes respectively. The first session was a training for the tasks. The second session was the main part of the experiment with 120 series of two trials. The third session included an urn task, and a final part with only perceptual choices and confidence ratings. Each task is detailed in the following subsections.

Perceptual Task (Type 1)

The perceptual task was a color numerosity task. On each perceptual trial, an array of 10×20 colored squares was presented for 1 second. Each square was either brown, purple, blue or green. Participants had to indicate which of these 4 colors was most present over the whole array. The

dominant color was determined randomly on each trial, with equal probability for each color.

The difficulty of the task was determined by the proportion of the dominant color relative to the other colors, and calibrated for each participant during the first session using the accelerated stochastic approximation procedure (Kesten, 1958). We aimed at 50% accuracy (recall that chance level is 25% accuracy in this task), in order to maximize the number of series for which there would be 1 correct and 1 incorrect answer. During the second session, task difficulty was constant at the level defined based on the first session.

Additional metacognitive task: Type 2 confidence ratings

In Experiment 2 only, after each perceptual decision, participants had to give their confidence on a scale ranging from 25% (i.e. chance level) to 100%. To ensure that participants did not confuse the Type 2 and Type 3 confidence tasks, the two scales did not have the same orientation (horizontal for Type 2, vertical for Type 3) and color (for Type 2, the confidence scale was colored according to the perceptual response, for Type 3 the scale was always gray). The difference was also emphasized in the instructions.

Forced choice (Type 2) and meta-metacognition (Type 3)

After a pair of perceptual trials, participants learned whether they have made 0, 1 or 2 correct perceptual decisions in these two trials. When both decisions were correct, or when both were incorrect, participants moved on to the next pair. Only in cases where exactly 1 perceptual decision was correct and 1 was incorrect, they had to indicate whether the first or the second was the correct one. The percentage of series for which participants had 1 correct and 1 incorrect answer was around 50% ($M_{freq} = 47.91\%$, $SD_{freq} = 5.02$ for Experiment 1, $M_{freq} = 47.43\%$, $SD_{freq} = 4.11$ for Experiment 2).
Finally, participants had to give their confidence in this Type 2 forced choice ("What are the chances that your choice is correct?"), on a vertical scale ranging from 50% to 100%. Participants were informed that 50% was the chance level in the Type 2 choice task.

Urn task

In the third session of Experiment 2, the urn task was introduced to participants. Participants were presented with two urns, filled with green and red balls. Participants were told that one ball had been randomly drawn in each urn, resulting in a draw of 1 green ball and 1 red ball. They were then asked to assess the likelihood that the green ball came from the first urn. Critically, if one considers that the proportion of green ball in an urn correspond to the likelihood that a given perceptual decision is correct, this task is formally equivalent to Type 3 judgments in the main session.

From trial to trial, the question was always the same but the proportions of green balls in each urn differed. In fact, those proportions corresponded to the Type 2 confidences reported by participants in the second session (the day before). More precisely, the proportion of green balls in one urn corresponded to the highest Type 2 confidence and the proportion of green balls in the other urn

to the lowest Type 2 confidence. This choice of not preserving the Type 2 confidences' order when presenting the urns was made to simplify the urn task for participants: they knew that the targeted urn was always the same (i.e. the one on the left side of the screen) and that the probability they had to report could never be smaller than 50%. The number of trials in this task depended on the second session and was equal to the number of series for which the participants had 1 correct and 1 incorrect answer. A practical example is presented in Figure 3.

Participants had a quick training of 10 trials for this task, in which the proportions of green balls in each urns were randomly drawn and the correct answer was provided to participants.

Measuring metacognition without Type 3 evaluations

Finally, after completing the urn task in Experiment 2, participants had to do 240 trials of the perceptual task with a Type 2 confidence rating after each trial. In this part, there were no Type 2 forced choice nor Type 3 evaluations. The goal of this final part was to ensure that Type 3 evaluations would not impact Type 2 ratings.

Payment and incentives

Participants' answers to the three tasks (perceptual, Type 2 choices and Type 3 ratings) were monetarily incentivized. At the end of the experiment, one pair of trials was randomly chosen for payment. Participants received 6 euros for each correct perceptual decision in this selected pair. In addition, if only 1 of the 2 answers in the selected pair of trials was correct, participants could still receive the 6 euros for the perceptual decision. Specifically, the participant was offered an exchange between their Type 2 response and a lottery ticket with an unknown probability P of success, generated randomly by computer. If P is greater than the Type 3 confidence, then the participant's bonus is determined by the lottery. If not, the bonus is determined by the accuracy of the Type 2 response. This procedure is inspired from the Becker-DeGroot-Marschak mechanism (Becker et al. (1964)), and from prior application to perceptual confidence (Massoni et al. (2014)). Here, this mechanism was adapted to incentivize both the metacognitive and meta-metacognitive tasks. It was presented to participants as a way to maximize their earning by providing accurate Type 2 decisions and Type 3 confidence ratings. Instructions and a training phase with feedback (30 series in the first session) were included to make sure that participants understood the mechanism.

In addition in Experiment 2, we incentivized Type 2 confidence ratings in the second and third sessions. To do so, one trial was drawn in each of these sessions, and subjected to a probability matching mechanism now applied to the Type 1 and Type 2 judgments. Participants could get an additional 4 euros for each of those trials, depending on the outcome of the mechanism.

For the urn task in Experiment 2, performance was not incentivized so as to avoid participants from being tempted to compute the likelihood using an online calculator. They received a fixed payment of 4 euros for this task.

Measures: Type 3 bias and sensitivity

We define Type 3 bias as the difference between average Type 3 confidence and Type 2 performance (the percentage of correct choices in the Type 2 task). We define Type 3 sensitivity as the area under the Type 3 roc curve (AUROC3). This measure directly translates the Type 2 AUROC measure at the metacognitive level to the meta-metacognitive level. More specifically, we use each Type 3 confidence level as a criterion that splits the data between high and low Type 3 confidence. Then, for each split of the data, we calculated the proportion of high Type 3 confidence trials among correct Type 2 forced choice decisions (Type 3 hit rate) and the proportion of high Type 3 confidence trials among incorrect forced choice decisions (Type 3 false alarm rate). Plotting these two values across all possible split produces a Type 3 ROC curve. The area under this Type 3 ROC curve was used as a measures of meta-metacognitive sensivity.

Measures: pseudo-Type 3 rating calculated from the urn task

Each trial in the urn task corresponded to a pair of trial in the second session, and allowed us to define a pseudo-Type 3 rating (noted $\widetilde{c}_3$) for that pair. We denote by c_A and and c_B the proportion of green balls in the two urns. From Bayes rule, the probability that the green ball comes from urn A and the red ball comes from urn B is as follows:

$$p^* = \frac{c_A \times (1 - c_B)}{c_A \times (1 - c_B) + c_B \times (1 - c_A)}$$

Since c_A and c_B were in fact the highest and lowest Type 2 confidence in the corresponding pair of trial in the second session, p^* is the (Bayesian) probability that the correct trial was the one with the highest Type 2 confidence. It thus corresponds to a Type 3 evaluation. We note p^{sub} participants' report in the urn task, which corresponds to their subjective evaluation of p^*. We can think of p^{sub} as the subjective aggregate (i.e. not necessarily Bayesian) of all the information contained in Type 2 ratings about the accuracy of the highest confidence trial. In addition, because participants' actual Type 3 judgments referred to the selected trial in the Type 2 forced choice (which was not always the trial with the highest confidence rating as documented in the Results), we had to align the pseudo-Type 3 according to the selected trial. To do so, $\widetilde{c}_3$ was set equal to p^{sub} when the participant chose the highest Type 2 confidence trial in the forced choice decision, and equal to $1 - p^{sub}$ otherwise.

Measures: Leveraging information to improve Type 2 choices

Our objective was to assess how Type 2 and Type 3 ratings could be used to improve the accuracy of Type 2 forced choices. To achieve this, we analyzed various simulated agents, each equipped with different types of information (i.e. different predictors) for making Type 2 forced choices. Then, for each of these agents, we searched for the linear relationship of predictors that would maximize the accuracy of their Type 2 forced choices.

More specifically, our approach consisted of two main steps. Firstly, we built three predictors (x_1, x_2 and x_3) to represent the likelihood that the first trial is correct, based on three different

sources of information. The first predictor x_1 conveys the information contained in the participant's Type 3 confidence rating: it is equal to the participant's Type 3 rating when the first trial was the one selected in the Type 2 forced choice, and to 1 minus this Type 3 rating otherwise. The second predictor x_2 is derived from participants' answer in the urn task: it is equal the pseudo-Type 3 rating when the first trial was the one selected in the Type 2 forced choice, and to 1 minus this pseudo-Type 3 rating otherwise. Finally, the third predictor x_3 reflects the comparison between participants' Type 2 ratings for the two trials of the pair: it is equal to 1 if the first trial was associated with the highest Type 2 rating, to .5 if both trials received the same Type 2 rating and to 0 if the first trial was associated with the lowest Type 2 rating.

Secondly, these different predictors were used to define Type 2 choices in 3 simulated agents, as follows:

$$\text{Type 2 forced choice} = \begin{cases} \text{Trial 1} & \text{if } \alpha x_i + \beta \geq .5 \\ \text{Trial 2} & \text{if } \alpha x_i + \beta < 0.5 \end{cases}$$

where α and β are free parameters (defined separately for each participant), such that the simulated agent would maximize the accuracy of its Type 2 forced choices. Similarly, we defined a full agent which used all three predictors to maximize its Type 2 accuracy, using the following decision rule:

$$\text{Type 2 forced choice} = \begin{cases} \text{Trial 1} & \text{if } \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \beta \geq .5 \\ \text{Trial 2} & \text{if } \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \beta < .5 \end{cases}$$

References

Becker, G. M., DeGroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral science*, 9(3):226–232.

Boldt, A. and Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *Journal of Neuroscience*, 35(8):3478–3484.

Buratti, S., Allwood, C. M., and Kleitman, S. (2013). First-and second-order metacognitive judgments of semantic memory reports: The influence of personality traits and cognitive styles. *Metacognition and learning*, 8:79–102.

Charles, L., Van Opstal, F., Marti, S., and Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *Neuroimage*, 73:80–94.

Daniel, R. and Pollmann, S. (2012). Striatal activations signal prediction errors on confidence in the absence of external feedback. *Neuroimage*, 59(4):3457–3467.

Dunlosky, J., Serra, M. J., Matvey, G., and Rawson, K. A. (2005). Second-order judgments about judgments of learning. *The Journal of general psychology*, 132(4):335–346.

Elosegi, P., Rahnev, D., and Soto, D. (2023). Think twice: Re-assessing confidence improves visual metacognition.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., and Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, 4(6):385–390.

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

Hainguerlot, M., Vergnaud, J.-C., and de Gardelle, V. (2018). Metacognitive ability predicts learning cue-stimulus associations in the absence of external feedback. *Scientific Reports*, 8(1):5602.

Holroyd, C. B. and Coles, M. G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological review*, 109(4):679.

Kesten, H. (1958). Accelerated stochastic approximation. *The Annals of Mathematical Statistics*, pages 41–59.

Konishi, M., Berberian, B., de Gardelle, V., and Sackur, J. (2021). Multitasking costs on metacognition in a triple-task paradigm. *Psychonomic Bulletin & Review*, 28(6):2075–2084.

Lee, A. L., Ruby, E., Giles, N., and Lau, H. (2018). Cross-domain association in metacognitive efficiency depends on first-order task types. *Frontiers in Psychology*, 9:2464.

Mamassian, P. and de Gardelle, V. (2022). Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychological Review*, 129(5):976.

Maniscalco, B. and Lau, H. (2015). Manipulation of working memory contents selectively impairs metacognitive sensitivity in a concurrent visual discrimination task. *Neuroscience of Consciousness*, 2015(1):niv002.

Massoni, S., Gajdos, T., and Vergnaud, J.-C. (2014). Confidence measurement in the light of signal detection theory. *Frontiers in psychology*, 5:1455.

Mazancieux, A., Fleming, S. M., Souchay, C., and Moulin, C. J. (2020). Is there a g factor for metacognition? correlations in retrospective metacognitive sensitivity across tasks. *Journal of Experimental Psychology: General*, 149(9):1788.

Nelson, T. O. and Narens, L. (1994). Why investigate metacognition. *Metacognition: Knowing about knowing*, 13:1–25.

Pleskac, T. J. and Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological review*, 117(3):864.

Rabbitt, P. A. (1966). Errors and error correction in choice-response tasks. *Journal of experimental psychology*, 71(2):264.

Rahnev, D. and Fleming, S. M. (2019). How experimental procedures influence estimates of metacognitive ability. *Neuroscience of consciousness*, 2019(1):niz009.

Recht, S., de Gardelle, V., and Mamassian, P. (2021). Metacognitive blindness in temporal selection during the deployment of spatial attention. *Cognition*, 216:104864.

Recht, S., Jovanovic, L., Mamassian, P., and Balsdon, T. (2022). Confidence at the limits of human nested cognition. *Neuroscience of consciousness*, 2022(1):niac014.

Recht, S., Mamassian, P., and De Gardelle, V. (2019). Temporal attention causes systematic biases in visual confidence. *Scientific Reports*, 9(1):11622.

Reyes, G., Silva, J. R., Jaramillo, K., Rehbein, L., and Sackur, J. (2015). Self-knowledge dim-out: stress impairs metacognitive accuracy. *PLoS One*, 10(8):e0132320.

Reyes, G., Vivanco-Carlevari, A., Medina, F., Manosalva, C., De Gardelle, V., Sackur, J., and Silva, J. R. (2020). Hydrocortisone decreases metacognitive efficiency independent of perceived stress. *Scientific Reports*, 10(1):14100.

Sherman, M. T. and Seth, A. K. (2024). Knowing that you know that you know? an extreme-confidence heuristic can lead to above-chance discrimination of metacognitive performance. *Neuroscience of Consciousness*, 2024(1):niae020.

Ye, Q., Zou, F., Dayan, M., Lau, H., Hu, Y., and Kwok, S. C. (2019). Individual susceptibility to tms affirms the precuneal role in meta-memory upon recollection. *Brain Structure and Function*, 224:2407–2419.

Yeung, N. and 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., and Rahnev, D. (2023). Common computations for metacognition and meta-metacognition. *Neuroscience of Consciousness*, 2023(1):niad023.

Supplementary Materials
Confidence in metacognition
Type 3 judgments in the context of perceptual decisions

1 Regression Analyses

		Model 1	*Model 2*	*Model 3*
Intercept	Estimate (Std)	1.01 (0.07)	0.99 (0.07)	0.99 (0.07)
	z	13.45	13.52	13.53
	p	$< .001$	$< .001$	
Type 3	Estimate (Std)	0.31 (0.05)	0.39 (0.05)	
	z	5.84	7.84	✗
	p	$< .001$	$< .001$	
Pseudo-Type 3	Estimate (Std)	0.32 (0.06)		0.41 (0.05)
	z	5.83	✗	7.82
	p	$< .001$		$< .001$
	AIC	2254	2285	2286
	N_{Obs}	1974	1974	1974
	N_{Sub}	37	37	37

Table 1: Type 3 and Pseudo-Type 3 are z-scored, within each participant. Model 1 includes both predictors. Model 2 only includes Type 3 and Model 3 Pseudo-Type 3. Each model includes a random intercept at the participant level.

2 Supplementary experiment without feedback after each series

Design and Participants

Experiment 3 follows a similar design than Experiment 2, with one modification in the procedure: participants engage in both the Type 2 forced-choice task and the Type 3 confidence task following each pair of trials, without receiving feedback regarding the number of correct answers in the pair. In practice, after every two perceptual trials, participants are asked to identify which trial they believe is most likely to be correct and then they are asked to express their confidence in this decision, in the event that only one of their two responses is accurate. Specifically, they are asked, "You have

chosen answer 1 (or 2). In the event that only one of your answers is correct, what are the chances that you chose the correct answer?".

This modification in the experimental design enables an investigation into whether the outcomes observed in Experiment 2 are influenced by participants receiving feedback on the number of correct answers in each pair. The subsequent section provides a concise overview of the findings from Experiment 3. Analyses are carried out specifically on the subset of trial pairs where participants provided one correct and one incorrect answer, ensuring that those analyses are comparable with those carried out in Experiment 2.

In Experiment 3, we recruited 27 adults using the online database of the Parisian Experimental Economics Laboratory (LEEP) and we removed data from 4 participants because their performance at the Type 2 task was not higher than the chance level (one sided binomial test against 50%). Our analyses are therefore based on the data of 23 participants.

Results

First, we replicated the basic findings of Experiment 1 and 2: Type 2 accuracy was higher when Type 3 confidence was high than when it was low ($M_{high} = 78.07$, $M_{low} = 67.50$, $t(22) = 3.32$, $p = .003$), Type 3 bias indicated that individuals overestimated their ability to identify the correct trial in the pair (Type 3 confidence: 85.66% vs. Type 2 accuracy: 72.76% ; $t(22) = 5.76$, $p < .001$; Figure 1A) and Type 3 sensitivity was greater than chance ($M_{sensitivity} = 62.61\%$, $t(22) = 5.10$, $p < .001$; Figure Figure 1B). When comparing Experiment 2 and 3, we found no differences in terms of Type 1 perceptual performance, Type 2 forced choice accuracy, Type 2 sensitivity, Type 2 overconfidence, Type 3 sensitivity, and Type 3 overconfidence (all $p > .3$).

In the urn task, participants' reported probabilities were well aligned with the Bayesian probability (Figure 1C), with high correlations between the two measures ($Mean_r = .62$, $SD = 0.27$, $t(22) = 10.99$, $p < .001$). Reported probabilities were slightly higher than the Bayesian probabilities ($M_{Dif} = -3.48$, $t(22) = -3.141$, $p = .005$). Importantly, there was no significant differences between Experiment 2 and 3 regarding those two metrics ($Mean_r^{Exp2} = 0.71$, $Mean_r^{Exp3} = 0.62$, $t(58) = 1.21$, $p = .232$; $Mean_{Dif}^{Exp2} = -1.72$, $Mean_{Dif}^{Exp3} = -3.48$, $t(58) = 0.76$, $p = .450$). Type 3 confidences were only weakly correlated with pseudo-Type 3 ratings ($Mean_r = 0.24$, $SD = 0.21$, $t(22) = 5.60$, $p < .001$; Figure 1D), and on average, Type 3 confidences were significantly higher than pseudo-Type 3 ratings ($M_{Type3} = 85.69\%$ vs. $M_{Pseudo-Type3} = 66.17\%$, $t(22) = 9.24$, $p < .001$). Again, no significant differences were found between Experiment 2 and 3 on these metrics ($Mean_r^{Exp2} = 0.34$, $Mean_r^{Exp3} = 0.24$, $t(58) = 1.86$, $p = .068$; $Mean_{Dif}^{Exp2} = 20.59$, $Mean_{Dif}^{Exp3} = 19.42$, $t(58) = 0.41$, $p = .680$).

We then compared regression models in which Type 2 forced-choice accuracy was predicted from both Type 3 confidence and pseudo-Type 3 ratings, or from only one of these variables. Similar to Experiment 2, the best-fitting model ($AIC = 1351$) included a fixed intercept, a fixed main effect for each parameter with no interaction, and a random intercept at the participant level. In this analysis as well, Type 2 accuracy was significantly predicted by both Type 3 confidence ($z = 5.49$, $p < .001$) and pseudo-Type 3 ratings ($z = 6.32$, $p < .001$).

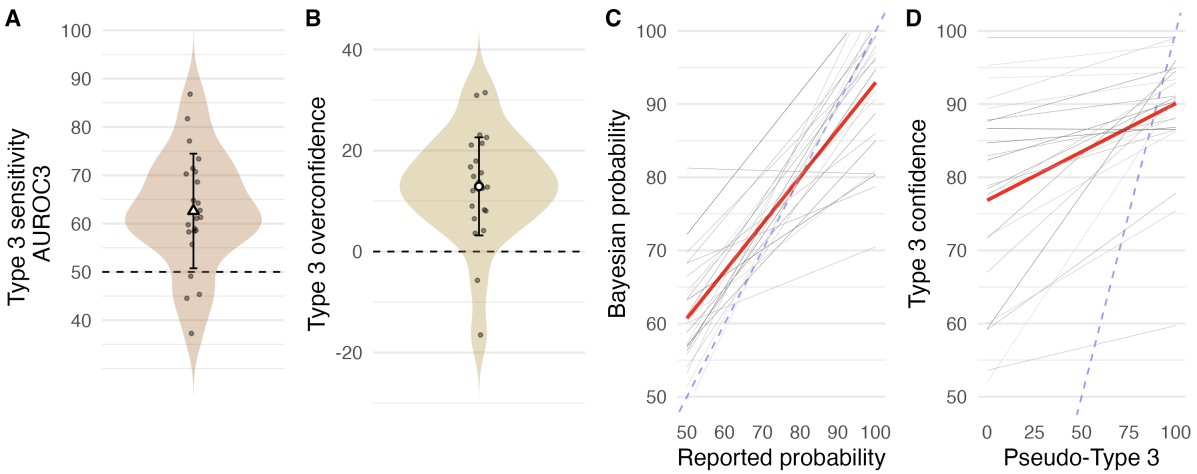

Figure 1: **A**. Distribution of Type 3 sensitivity. **B**. Distribution of Type 3 overconfidence. **C**. Bayesian probability plotted against the reported probability in urn task. **D**. Type 3 confidence plotted against Pseudo-Type 3. *Note*: Analyses carried out on the subset of trial pairs where participants provided one correct and one incorrect answer.

Finally, we observed similar rates of inconsistent Type 2 forced choices in Experiment 3 compared to Experiment 2 ($M^{Exp2} = 23.34$, $M^{Exp3} = 25.09$, $t(58) = -0.59$, $p = .560$) and similar accuracy for those inconsistent choices ($M^{Exp2} = 57.13$, $M^{Exp3} = 54.39$, $t(58) = 0.51$, $p = .614$). Participants' accuracy for inconsistent decisions was also above chance in Experiment 3 ($M_{accuracy} = 54.39$, $Chance = 50$, $t(22) = 1.06$, $p = .302$; when weighted by the number of inconsistent decisions, which varied across participants: $M_{accuracy} = 56.25$, $t(22) = 2.24$, $p = .035$).