

Dynamic sources of evidence supporting confidence judgments and error detection

Authors: Lucie CHARLES^{a,b}, Nick YEUNG^a

Affiliations:

^a Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, United Kingdom

^b Institute of Cognitive Neuroscience, University College London, London WC1N 3AR United Kingdom

Corresponding author: Lucie Charles

Institute of Cognitive Neuroscience Alexandra House, 17 Queen Square WC1N3AR London

Tel: +44 1865 271302

lucie.charles.ens@googlemail.com

Number of pages: 29

Number of figures: 6

Number of words for abstract: 145

Number of words for Introduction: 1429

Number of words for Methods: 1065

Number of words for Results: 2322

Number of words for Discussion: 1788

Number of words Main Text: 6604

Number of References: 42

Abstract

Our decisions are accompanied by a subjective sense of confidence about whether the choices we have made are correct or erroneous. Here we investigate the information on which these confidence judgments are based, and how they relate to the decision itself, by studying how fluctuations in perceptual information influence decisions and second-order metacognitive evaluations of confidence and accuracy. Human participants judged which of two dynamically changing stimuli contained more dots, under instructions emphasizing either speed or accuracy. Crucially, stimuli remained visible after the decision, before participants rated their confidence in their choice. We found that confidence and error detection depended on the balance of stimulus evidence accumulated in the periods both preceding and following the initial decision, regardless of speed-accuracy instruction. These findings suggest a shared computational basis for error detection and confidence judgments, with implications for current models of metacognitive evaluation of decision processes.

Keywords: Metacognition, Error-detection, Confidence, Decision-making, Reverse correlation

Significance Statement

- We investigated how fluctuations in incoming information impact decisions, changes of mind and levels of confidence in human perceptual decisions.
- We show that confidence and error detection depend on evidence accumulated both before and after the decision, regardless of the speed imposed to the decision
- Our findings suggest a shared computational basis for confidence judgment, error detection and changes of mind.

1. Introduction

The ability to evaluate and revise decisions is a core function in adaptive behavior. There is thus considerable interest in the mechanisms supporting metacognitive evaluations of decision processes (Fleming & Frith, 2014; Peters et al., 2017; Resulaj, Kiani, Wolpert, & Shadlen, 2009; van den Berg, Zylberberg, Kiani, Shadlen, & Wolpert, 2016), to understand how humans and other animals detect their errors and represent and act upon graded judgments of confidence in their choices (Kepecs & Mainen, 2012; Kepecs & Mainen, 2014). The present study investigates confidence and error detection, and the relationship between them, in terms of their sensitivity to dynamics of stimulus evidence during the decision process.

Although confidence judgments and error detection are conceptually related, research on these functions has historically developed separately, reflected in differences in methodology and theoretical emphasis in the respective fields (Yeung & Summerfield, 2012, 2014). Thus, on one hand, the cognitive and neural correlates of error-detection have been studied for many years using tasks in which the decision itself is trivial—e.g., judging whether a centrally presented arrow stimulus points left or right, or whether a digit is greater or less than 5—and errors are induced through pressure to respond quickly (Gehring, Goss, Coles, Meyer, & Donchin, 1993; Rabbitt, 1966). Dominant theoretical accounts propose that the resulting “fast guess” errors are detected by continued processing of the stimulus after the initial response, such that “a more accurate consensus will accumulate and the earlier mistake will become apparent” (Rabbitt & Vyas, 1981). Most or all current theories of error detection share this core assumption, and differ primarily in terms of precisely how post-decisional evidence might be harnessed to support error detection (Yeung & Summerfield, 2012).

By contrast, studies of metacognitive judgments of confidence have typically used stimulus ambiguity rather than time-pressure as a source of errors, e.g., asking participants to identify the longer of two lines of very similar length (Henmon, 1911), or to detect a Gabor patch of slightly greater contrast than frequent standard stimuli (Fleming, Weil, Nagy, Dolan, & Rees, 2010), before giving a graded rating of confidence in their initial choice. In such experiments, participants often remain unsure whether they responded correctly or incorrectly even when judgments are unspeeded, and are typically asked to judge their confidence on a scale ranging from feeling that they are guessing to feeling certain they are correct (i.e., with no option to indicate explicit error detection) (Fleming et al., 2015, 2010; Maniscalco & Lau, 2012).

Influential early theories of confidence correspondingly did not allow for changes of mind and error detection (which depend on post-decisional processing), instead proposing that confidence reflects features of the decision process up to the time of the decision, such as the balance of evidence accumulated for competing response options (Vickers & Packer, 1982; Vickers, 2001) or the time taken to reach the decision (Audley, 1960). According to these *decision-locus* models (and more recent variants, Kiani, Corthell, & Shadlen, 2014a, Kepecs & Mainen, 2014, Zylberberg, Barttfeld, Sigman, & Pereira, 2012), confidence depends critically on the strength and consistency of evidence accumulated up to the time of the decision.

However, recent evidence indicates that decision confidence, like error detection, depends critically on continued processing of available evidence even after an initial decision is made. For example, the *resolution* of confidence judgments—the degree to which subjective confidence predicts objective accuracy—is improved when greater time is allowed between initial choice and subsequent confidence judgment (Moran, Teodorescu, & Usher, 2015; Pleskac & Busemeyer, 2010; Resulaj et al., 2009; van den Berg, Anandalingam, et al.,

2016; Yu, Pleskac, & Zeigenfuse, 2015). One recent study proposed that this could be due to integration into confidence judgments of late evidence that is processed between internal commitment to a choice and overt expression of the decision (van den Berg, Anandalingam, et al., 2016). This study found that confidence varies according to stochastic fluctuations in evidence presented immediately preceding the overt response. Extending this idea, another recent study (Moran et al., 2015) found that, when possible, people continue to accrue perceptual evidence presented after their decision to inform their confidence judgments, confirming the crucial role of post-decisional process on confidence. Meanwhile, EEG studies suggest that confidence and error judgments are reflected in common neural signatures (in particular the Pe component) that unfold in the period after response (Boldt & Yeung, 2015; Murphy, Robertson, Harty, & O'Connell, 2015). Formal models of this post-decision accumulation process suggest that these *post-decisional locus* models may provide an integrated account of confidence and error judgments (Moran, Teodorescu, & Usher, 2015; Pleskac & Busemeyer, 2010; Yu, Pleskac, & Zeigenfuse, 2015).

This convergence notwithstanding, several open questions remain, which form the basis for the present study. Although it seems uncontroversial that a decision maker should take advantage of additional information presented after initial choice to form a more accurate confidence judgment and possibly revise that decision, there has been little systematic investigation of the distinct influences of pre- and post-decisional information on confidence and error detection. In particular, it remains unclear what impact pre-decisional evidence has compared to post-decisional evidence on evaluations of decisions. Interestingly, while some *pre-decisional locus* models predict that confidence should reflect the evidence available at the time of the response, *post-decisional locus* models have proposed various ways by which evidence continues to be accumulated to inform confidence (Moran, Teodorescu, & Usher, 2015, Collapsing Confidence Boundary model; Pleskac & Busemeyer,

2010, Two-stage Dynamic Signal Detection model; Van Den Berg et al., 2016). Going beyond *locus* models that focus only on the evidence available at a single point in time, our first goal was to investigate the timecourse over which decision evidence influences confidence judgments, error detection, and choice, both in the period leading up to the moment of the choice and in the post-decision period leading up to the moment of the confidence judgment.

Our second goal was to determine how speed-accuracy tradeoffs affect these dynamics of evidence accumulation for confidence judgments and error detection. It remains to be established whether speeded or self-paced decisions rely to the same extent on pre vs. post decisional evidence—a salient question given the differing emphasis on speeded vs. unspeeded tasks in studies of error detection and confidence, respectively (Yeung & Summerfield, 2012), and evidence that the speed-accuracy trade-off affects confidence and error detection (Baranski & Petrusic, 1998; Gehring et al., 1993). Although some earlier studies have investigated this question and found indirect evidence that speed/accuracy tradeoff might alter the balance between pre- and post-decisional evidence (Moran et al., 2015), no direct analysis has explored the time-course of this effect.

Our final goal was to investigate further the relationship between confidence and explicit error detection judgments. Only a few studies to date have investigated explicit error detection and confidence in the same experiment (Baranski & Petrusic, 1994; Boldt & Yeung, 2015; Scheffers & Coles, 2000; Yu et al., 2015). It therefore remains to be tested how these two judgments relate and whether they rely on corresponding dynamics of stimulus evidence.

To answer these questions, we present a systematic investigation of the decision dynamics supporting error detection, changes of mind and confidence judgments in a perceptual judgment task. Participants judged which of two dynamically changing stimuli

contained more dots on average over time (Figure 1), responding under instructions emphasizing either speed or accuracy, before providing a graded confidence judgment about the initial decision. Crucially, we directly retrieved what in the stimulus dynamics led participants to commit errors and detect them, using reverse correlation methods (Kiani & Shadlen, 2009; van den Berg, Anandalingam, et al., 2016; Zylberberg et al., 2012) to probe how stochastic signal fluctuations in the stimuli influenced participants' decisions and subsequent metacognitive evaluations. Importantly, in our design, stimuli continued to be visible after participants made their decision. Therefore, participants could continue sampling information before indicating their confidence, enabling us to determine how much they relied on post-decision evidence to evaluate the accuracy of their initial decision. Manipulation of time pressure induced participants to make errors, some of which remained undetected, allowing us to contrast dynamics of stimulus evidence associated with decisions that were or were not followed by a change of mind (ChoM), providing further insight into the mechanisms of error detection and confidence judgments.

2. Methods

1.1. Participants

Twenty-three right-handed participants, with normal or corrected-to-normal vision, gave informed consent to participate in the experiment (mean age 25.1 years). As our analysis focused on subjective confidence reports, five participants were discarded for using the confidence scale in a discretized manner in which they used the end-points or the middle of the scale in more than 30% of the whole of the trials, which left insufficient number of trials

to perform key analyses. This left 18 participants (11 female, mean age 25 years) in the final sample.

1.2.Task and Procedure

Stimuli were presented on a 20 inch CRT (Trinitron, Dell) monitor with a 60 Hz refresh rate using the MATLAB toolbox Psychtoolbox3. Stimuli were 4.7 cm large, resulting in a visual angle of 4.48 degrees when viewed from 60 cm, and were placed 3.9 cm (3.6 degrees) to the left and right of the fixation cross.

The trial started with a small increase in the size of the fixation cross (100 ms duration) reminding the participant to fixate the centre of the screen. After 100 ms, two empty grey boxes then appeared on the screen signalling the beginning of the trial. After 300 ms, dots appeared at random positions within the two grey boxes. The display was then updated every 50 ms, with the dots in each box changing randomly and independently in position and number.

The two boxes were 20-by-20 resolution, thus containing at most 400 dots. The number of dots presented at each time-sample in each box was drawn from a two Normal distributions respectively around either a high (212 dots) or low (188 dots) mean value and variance of 40 dots for both.

Participants' task was to identify the box with the higher mean. Participants responded with a left or right mouse-button click corresponding to the box they judged to contain more dots. Importantly, the boxes continued to be displayed for 1000 ms after the participant's response, again updated every 50 ms, allowing the participant to continue sampling the information after their initial decision.

Two types of blocks varied in time-pressure. In Speed blocks, participants were required to respond in less than 800 ms, whereas in Accuracy blocks the instructions

emphasised the importance of accurate responding within a correspondingly lenient response deadline of 3000 ms. In both block types, the time left to respond was indicated by a bar on the top of the screen that gradually filled up at each time sample.

At the end of each trial, after the dot stimulus display disappeared (always 1000 ms after their perceptual decision), participants were asked to indicate how confident they were in their preceding decision by moving a cursor along a 51-point scale ranging from “Sure Error” (scored as 0 on the scale) to “Sure Correct” (scored as 50 on the scale). Additionally, the word “Guess” was displayed in the middle (scored as 25 on the scale), which also corresponded to the initial position of the response cursor (to ensure that it would take the same time to move the cursor to each end of the scale, equating the effort and time to signal an error and a correct response). No time pressure was imposed for the confidence response. After participants had registered their confidence rating, the next trial started after a blank screen interval of 500 ms.

Participants completed 12 blocks of 40 trials. There were 6 blocks with each speed-accuracy instruction, randomly intermixed. Altogether the experiment lasted approximately 60 minutes.

1.3. Analysis of pre- and post-decisional evidence

Our first analysis quantified the evidence participants saw when making correct vs. erroneous decisions and when subjectively evaluating these decisions as correct or incorrect. For each trial, we extracted the number of dots at each time sample in each box. This allowed us to retrieve for each trial the distributions of dot numbers in each box over time, for two time periods of interest: from the onset of the stimulus to the response (i.e., the pre-decisional

interval) and from the response to the disappearance of the stimulus 1000 ms later (i.e., the post-decisional interval). To quantify the strength of the evidence for these two time periods on each trial, we used a standardized approach to measure the degree of overlap between the two distributions, calculating the Receiver Operating Characteristic curve and associated Area Under Curve (AUC). These measures allow us to quantify, for each trial, the degree of discriminability of the box with the higher mean number of dots compared to the one with the lower mean number in each of these time-periods. The advantage of this approach compared to simply calculating the average difference in the number of dots is that it provides a standardized measure of discriminability that is also sensitive to stimulus variability, thus quantifying the impact of the stimulus fluctuations on task difficulty. AUC varies between 0 and 1, with 1 indicating that the evidence unambiguously favored the objectively correct response, the midpoint (0.5) indicating that the evidence was perfectly ambiguous (i.e., noise fluctuations on average perfectly cancelled out the underlying difference in mean dot numbers between the two boxes), and values below 0.5 indicating that objective evidence favoured the alternative response (i.e., noise fluctuations were sufficiently large to outweigh the underlying mean difference).

The obtained trial-by-trial AUC values were then averaged together for each participant according to the conditions corresponding to the factorial combination of block type (Speed vs. Accuracy), time interval (pre- vs. post-decisional), objective decision accuracy (correct vs. error) and subsequent metacognitive evaluation (no change of mind vs. change of mind). Some participants had too few changes of mind to compute AUC values across the full factorial design—this was true for 1 participant in Speed blocks and 5 participants in Accuracy blocks. To maximize power, we therefore performed a three-way repeated measure ANOVA separately for Speed and Accuracy blocks, with factors time interval, decision accuracy, and metacognitive evaluation, allowing us to retain 17

participants for the former condition and 13 in the latter. Effect sizes were computed using η_p^2 measure for ANOVAs, while pairwise Cohen's d measure (d_z , Cohen, 1988) was computed for additional t-tests.

1.4.Reverse Correlation Analysis

To assess the dynamics of stimulus evidence predicting error detection and confidence judgments, our next analysis used a reverse correlation approach (Resulaj et al., 2009; Zylberberg et al., 2012). This method correlates observed behavior with the momentary, stochastic fluctuations in evidence that were built into our dot display stimuli. We separated trials according to block type, accuracy and change of mind, and retrieved for each trial the variation in the number of dots across time in each of the two boxes, the correct and the incorrect one (Figure 3, top row). We then normalized the two obtained time-courses by subtracting the mean number of dots in each box, respectively, and divided by the across time-sample variance (Figure 3, second row). The obtained trial-by-trial time-courses were then individually realigned to the onset of the motor response, and averaged separately for the low-mean and the high-mean box, then averaged across participants. To avoid averaging conditions with too few data points, we excluded from averaging data points that contained fewer than five trials and time-samples for which fewer than five participants had data. For display purposes, participants' individual time-courses were temporally smoothed by averaging together values of the two preceding and two following time-samples. Statistics were however computed on unsmoothed data.

To determine the moment at which the number of dots in the two boxes significantly deviated from each other, we computed between-participant statistics on the obtained averaged time-series using a cluster-based non-parametric test with Monte Carlo randomization (adapted from Maris and Oostenveld, 2007). This method allowed us to

identify clusters of time-points in which time-series of the two stimuli were significantly different while correcting for multiple comparisons (see Supplementary Material).

1.5. Statistical Power

The main analysis proposed in the present study investigated the time-course of stimulus fluctuations by means of within-participant non-parametric permutation statistics. This is a new analysis for which no empirical estimation of effect size is available in the literature. As such, the required sample size could not be estimated by means of a classical power analysis. Sample size was therefore chosen to be comparable to the previous studies most closely related to the present one (Van Den Berg et al., 2016, n=6; Zylberberg et al., 2012b, n=19). Nonetheless, we are able to use our secondary analysis of the Area-Under Curve (see Methods, below) to estimate the *a priori* statistical power of the analysis given our sample size. With a sample of 15 participants (the smallest number allowing for full factorial analysis with ANOVA on the AUC with time interval, accuracy and change of mind as within-participant factors), the smallest effect size detectable would be $\eta_p^2 = 0.11$ for ANOVA and Cohen-d = 0.78 for a two-tailed t-test (both taking alpha = 0.05 and power = 0.8 and assuming no correlation between repeated measures and no correction for non-sphericity, using G-Power software). The main effect of interest reported in the manuscript (3-way interaction between decision accuracy, time interval and presence of a change of mind on the AUC value) exceeded this limit, suggesting that our design had appropriate statistical power.

3. Results

1.6. Task performance

Participants judged which of two boxes contained on average more dots (Figure 1), with varying time-pressure. We first verified that our experimental manipulation of time-pressure affected participants' speed-accuracy trade-off as intended. Unsurprisingly, we found a significant difference in reaction times between Speed and Accuracy blocks (mean RTs of 638 ms vs. 1681 ms; Figure 2A, Cohen-d = -2.82, $t(17) = -12.0$, $p < 0.001$). Accuracy was significantly lower in blocks with Speed vs. Accuracy emphasis (Figure 2B, 68% vs. 80% correct, Cohen-d = -2.01, $t(17) = -8.54$, $p < 0.001$). Correspondingly, average confidence was lower in blocks with Speed vs. Accuracy emphasis (Figure 2C, 63% versus 68%, Cohen-d = -1.44, $t(17) = -6.13$, $p < 0.001$), showing that participants were able to monitor variations in their performance across conditions.

We next split the data according to trial-by-trial accuracy and investigated the use of the confidence scale separately for Error and Correct trials (Figure 2D-I). We discretized the confidence scale to separate trials that participants judged as correct (right-hand side of the confidence scale) from those judged as errors and characterized by a revision of the initial decision (i.e. "change of mind", ChoM, left-hand side of the scale). An ANOVA on these proportions revealed that, in line with previous research (Rabbitt, 1966), changes of mind occurred more frequently following error than correct responses ($F(1,17) = 110.35$, $p < 0.001$, $\eta_p^2 = 0.87$). Changes of mind were more frequent in Speed than in Accuracy blocks ($F(1,17) = 41.277$, $p < 0.001$, $\eta_p^2 = 0.71$). Furthermore, an interaction between accuracy and block type indicated that error detection rates were higher in Speed blocks than in Accuracy blocks,

whereas similar low rates of changes of mind were observed following correct responses across block types ($F(1,17) = 18.09, p < 0.001, \eta_p^2 = 0.51 \%$).

As a final observation, apparent in Figure 2 is that a sizeable proportion of confidence responses fell exactly at the midpoint of the confidence scale, corresponding to the guess response and the initial position of the cursor. The frequency distribution of responses on the confidence scale suggests that participants remained on this initial “guess” response for a range of low confidence responses that were therefore little used, indicating that a better methodology would have been to randomize the starting position of the confidence cursor. The proportion of guess responses was higher for errors than for correct trials ($F(1,17) = 11.67, p = 0.003, \eta_p^2 = 0.41$). No reliable difference between block types was observed ($F < 1$). As these “guess” trials could not be labelled as true changes of mind or perceived correct responses, we excluded them from further analysis.

1.7. Analysis of evidence available before and after the decision

The next analysis quantified the evidence that led participants to make correct responses, errors and, on occasion, to detect their errors. We used an area-under-the-curve (AUC) metric to quantify the degree to which the objectively-presented evidence favored the correct or incorrect decision across time-points and trials (with values greater than 0.5 indicating evidence favoring the correct decision, up to a maximum value of 1.0 where the evidence for this choice is perfectly unambiguous), how this evidence led to correct and erroneous decisions, and how it influenced the occurrence of changes of mind. The advantage of such an approach is that it allows quantification of objective stimulus discriminability using a standardized measure. Note, however, that results did not differ qualitatively when

performing the same analysis using the raw difference in dot number between the two boxes. AUC scores were averaged separately for evidence presented in the pre- vs. post-decisional period, and separately for each participant, for correct and error trials that were followed or not by a change of mind, before averaging across participants.

We found a main effect of decision accuracy on AUC in both block types (Speed blocks: $F(1,87)=17.08$ $p<0.05$, $\eta_p^2 = 0.63$; Accuracy blocks: $F(1,71)= 24.0$, $p<0.05$, $\eta_p^2 = 0.74$), indicating that participants made correct perceptual decisions when available evidence more strongly favored the correct response, and tended to err when the evidence was weaker. We also found an interaction between decision accuracy and time interval (Speed blocks: $F(1,87)=36.89$ $p< 10^{-3}$, $\eta_p^2 = 0.82$; Accuracy blocks: $F(1,71)= 24.46$, $p< 10^{-3}$, $\eta_p^2 = 0.70$), because this effect was of course restricted to evidence presented pre-decisionally. Notably, follow-up t-tests indicated that pre-decision AUC for all conditions was significantly larger than 0.5 (Figure 3, all $ps < 10^{-4}$). Thus, even for error trials, objective evidence available at the time of the decision favoured the correct response (Figure 3A-B).

Crucially, AUC scores also varied reliably as a function of whether participants changed their minds to indicate that an initial decision was incorrect, with a reliable main effect of changes of mind for Accuracy blocks ($F_{1,71}=10.21$, $p = 0.012$, $\eta_p^2 = 0.56$) and, for both block types, a reliable interaction between decision accuracy and changes of mind (Speed blocks: $F_{1,87}= 45.3$, $p = 0.001$, $\eta_p^2 = 0.82$ Accuracy blocks: $F_{1,71}=15.04$, $p = 0.005$, $\eta_p^2 = 0.65$) that further varied somewhat across time intervals (3-way interaction, Speed blocks: $F_{1,87}= 2.513$, $p=0.14$, $\eta_p^2 = 0.20$; Accuracy blocks: $F_{1,71}=7.892$, $p = 0.023$, $\eta_p^2 = 0.50$). The interaction between decision accuracy and changes of mind indicates that objective evidence in favour of the correct decision (i.e., increased AUC) had opposite effects on changes of mind as a function of initial accuracy: decreasing their likelihood after initially correct responses and increasing their likelihood after errors.

Importantly, follow-up analyses run separately for each time interval revealed that these effects were observed in both the pre- and post-decision AUC scores: Evidence in favour of the correct decision was stronger (i.e., AUC was higher) on trials in which participants detected their errors than for undetected errors, both in the pre- and post-decision periods for both Speed and Accuracy blocks (see Table 1). Meanwhile, evidence in favour of the correct decision was stronger when correct decisions were judged as such than when they were misjudged as errors, in the pre- decision period only for Speed emphasis blocks, and in the post-decision period for both block types (see Table 1).

Comparison AUC value	Block type	Time interval	t	p-value	Cohen-d
Error before ChoM > Error	Accuracy	Pre-resp	4.35	< 0.001	1.39
Error before ChoM > Error	Speed	Post-resp	3.05	0.006	0.92
Error before ChoM > Error	Accuracy	Pre-resp	4.06	< 0.001	0.95
Error before ChoM > Error	Speed	Post-resp	7.94	< 0.001	1.87
Correct > Correct before ChoM	Accuracy	Pre-resp	1.22	0.13	0.44
Correct > Correct before ChoM	Speed	Post-resp	2.14	0.03	0.64
Correct > Correct before ChoM	Accuracy	Pre-resp	1.83	0.04	0.54
Correct > Correct before ChoM	Speed	Post-resp	5.94	< 0.001	1.51

Table 1: Statistical results of the comparison of AUC values for errors followed or not by a change of mind (ChoM) and correct trials followed or not by a change of mind. Resp=response.

The effects just described relate to analyses including all trials. Control analyses confirmed that these effects were preserved in analyses excluding the subset of trials in which, due to noise fluctuations in dot numbers, the evidence presented up to the time of the response actually favored the incorrect response (i.e., trial-wise AUC < 0.5, see Figure S1). These trials occurred more frequently in Speed blocks (15% of trials vs. 3.5% in Accuracy blocks), and were associated with faster RTs than for other trials (Speed blocks: 614 vs. 640 ms, $t(13) = -3.56$, $p < 0.001$, Cohen-d = -0.84; Accuracy blocks : 1289 vs. 1678ms, $t(15) = -$

4.94, $p < 0.001$, Cohen- $d = -1.35$), indicating that these trials corresponded to fast guesses with shorter sampling-time of the stimulus display.

1.8.Reverse Correlation Analysis

Our next set of analyses focused on the dynamics of evidence accumulation influencing both initial decision and metacognitive evaluation. To this end, we ran a reverse correlation analysis to retrieve, for each time-point, the empirical kernels on which initial decisions and subsequent confidence judgments are based, according to across-trial averages of systematic biases in noise fluctuations across different trial subsets (Resulaj et al., 2009; Zylberberg et al., 2012). Figure 4 illustrates the logic of the analysis, showing stimulus-aligned averages of noise fluctuations in the low-mean box (i.e., the box the box with fewer dots on average, red lines) and high-mean box (i.e., the box with more dots on average, blue lines), for correct trials separately in Speed and Accuracy blocks. On correct trials, noise fluctuations in both boxes favored the ultimate choice. Thus, the average noise fluctuation was positive in the high-mean box (i.e., it contained more dots than its true already high mean) and negative in the low-mean box (i.e., it contained even fewer dots than its already low mean). The difference between boxes was significant from 50 – 700 ms in Speed blocks, and 50 – 850 ms in Accuracy blocks. In this way, the reverse correlation method identifies time periods in which stimulus evidence consistently influences participants' decisions across trials (cf. (Zylberberg et al., 2012). However, this stimulus-aligned analysis provides limited information about signal dynamics in relation to the time of the decision.

Our key analyses focused on trial-by-trial time-courses aligned to the onset of the response (Figure 5). Considering first the correct trials that were judged by participants as

correct (i.e., no change of mind), averaged evidence carried in noise fluctuations significantly deviated from the mean in the period from -650 ms to 0 ms before the onset of the response for Speed blocks (Figure 5A) and from -950 ms to -300 ms for Accuracy blocks (Figure 5B). These findings indicate that consistent fluctuations in evidence strength were only observed just prior to initiation of the motor response, whereas evidence presented at the earliest periods of stimulus processing did not systematically correlate with the decision reached. This pattern is necessarily observed in Speed blocks, in which a tight decision deadline was imposed, but the pattern was similarly evident in blocks with Accuracy emphasis where average RT exceeded 1500 ms.

Turning next to error trials that were not followed by a change of mind (i.e., errors that remained undetected), analysis of signal dynamics revealed the inverse pattern to the one observed in correct trials: On these trials, evidence preceding the response strongly favoured the incorrect decision, with the averaged noise fluctuation being reliably negative in the high-mean box (i.e., containing fewer dots than its true underlying mean) and positive in the low-mean box (i.e., containing more dots than its true underlying mean). These differences peaked at -500 ms and -700ms before the response for Speed (Figure 5G) and Accuracy emphasis blocks (Figure 5H), respectively. Thus, as with the evidence kernel observed on correct trials, deviations in noise fluctuations were observed in both the low- and high-mean boxes, suggesting sampling of both parts of the stimulus display. Interestingly, contrary to the pattern observed for correct trials without changes of mind, these differences persisted even after the decision for both types of blocks, albeit reaching significance only in Accuracy blocks (Figure 5H, 50 to 350 ms after response time-period). It appears that errors remained undetected only if noise fluctuations continued to favor the incorrect response after it was produced. Overall, therefore, these results show that errors that remain undetected are

characterized by noise fluctuations that vote against the correct response and continue to do so after the initial decision.

A different pattern was observed for errors followed by changes of mind. These trials were marked by evidence favouring the incorrect response in the early time-window before the response, an effect that was significant between -700ms and -400 ms relative to the response for Speed emphasis blocks (Figure 5E), and from -1200 to -500 ms for Accuracy blocks (Figure 5F). However, the pattern reversed around -300 ms before the response such that noise fluctuations began to favour the correct response. In Speed emphasis blocks, this effect was reliable from -200 ms before the response to 850 ms after it; in Accuracy blocks the effect was reliable from 100 to 400 ms after the response. The pattern suggests that while evidence presented before the response influenced the initial incorrect decision, evidence presented immediately before the response and for a sustained period afterwards continues to be accumulated that can lead to a change of mind about the decision and therefore detection that the initial response was incorrect. Interestingly, a similar pattern was observed in correct trials followed by a change of mind (Figure 5C-D), with dynamics of evidence accumulation also exhibiting a reversal in the direction of evidence regarding the choice. However, these trials occurred rarely (Figure S1) and the analysis was correspondingly underpowered, with the only statistically reliable effect being a brief period in Speed emphasis blocks (50 to 350 ms after the response) in which noise fluctuations favored the incorrect response.

1.9. Correlation between trial-by-trial balance of evidence and confidence before and after the response

Collectively, the reverse correlation results (Figure 5) indicate that error detection is influenced by evidence presented before and after the decision. However, these results do not indicate whether subtle variations in the level of confidence in correct decisions can likewise be explained in terms of fluctuations in evidence sampling, with particular interest in whether the ultimate confidence judgment is influenced by evidence presented after the initial choice. To investigate this issue, we computed the cumulative evidence for different levels of confidence observed specifically on correct trials without changes of mind (using cumulative evidence so that small differences in evidence are more apparent than in the moment-by-moment reverse correlation plots shown in Figure 5). For this analysis, we divided the correct-trial confidence distribution into quartile bins, then sorted trials into bins and averaged the cumulative evidence over time across trials within each bin, separately for the low- and high-mean boxes. The resulting curves showed systematic variations as a function of explicitly-reported confidence (insert Figure 6), with higher confidence observed as a function of higher cumulative evidence in the high-mean box and lower cumulative evidence in the low-mean box. These differences emerged in the pre-response period, but continued to develop well after the response into the post-decisional period, both for Speed and Accuracy emphasis blocks.

To confirm this result and test its significance, we computed the balance of evidence between the two boxes (i.e., the degree to which noise fluctuations on average favoured the correct vs. incorrect response) and regressed it against the ultimate confidence level. We performed these regressions separately for each participant for cumulative evidence from the time interval before the response and after it, to determine whether pre- and post-response evidence independently influenced the ultimate confidence judgment. We found significant positive correlations between evidence strength and confidence for both time intervals. This result was observed for both Speed and Accuracy blocks, indicating that it was not only an

482 effect of speed-accuracy trade-off that occurs when a tight response deadline is imposed
483 (Figure 6 main box plot, correlation coefficients across participants significantly greater than
484 0: Speed Block, Pre-response $t(17) = 6.8$ $p < 10^{-4}$; Speed Block, Post-response $t(17) = 3.82$ p
485 < 0.001 ; Accuracy Block, Pre-response $t(17) = 5.5$ $p < 10^{-4}$; Accuracy Block, Post-response
486 $t(17) = 3.33$ $p < 0.001$).

4. Discussion

The present study provides a systematic investigation of the way in which evidence—in terms of stochastic fluctuations in dynamically evolving stimuli—predicts the occurrence of errors, changes of mind, detection of errors, and graded levels of confidence in an initial decision. Our findings extend previous results (Murphy et al., 2015; van den Berg, Anandalingam, et al., 2016) in providing only partial support for current models of error processing and decision confidence. Thus, with some notable exceptions (e.g., Van Den Berg et al., 2016), extant models of decision confidence place emphasis on information available exclusively at the time of choice (e.g., Kiani, Corthell, & Shadlen, 2014a, Kepecs & Mainen, 2014, Zylberberg, Barttfeld, Sigman, & Pereira, 2012) or that is accumulated post-decisionally (Moran et al., 2015; Yu et al., 2015), but not both. Meanwhile, models of error processing focus almost exclusively on post-decisional processing as the basis for error detection (Yeung & Summerfield, 2014). Importantly, we found that confidence judgments and error detection are similarly influenced by the strength of the evidence presented both before and after the response, and that this dual influence was observed regardless of whether participants responded under speed or accuracy emphasis.

These findings shed new light on the mechanisms of error detection and confidence judgments. Early models of confidence were based on the intuition that confidence should reflect the strength of evidence supporting the initial decision (Audley, 1960; Festinger, 1943; Vickers & Packer, 1982). This assumption provides an elegant account of many empirically observed features of confidence judgments such as their dependence on task difficulty and response time (Kiani et al., 2014; Kiani & Shadlen, 2009; Vickers & Packer, 1982). More recently however, these decision locus models have been altered to capture the intuition that evaluation of a decision should be sensitive to continuing reflection even after an initial

choice, thus allowing for changes of mind (Moran et al., 2015; Pleskac & Busemeyer, 2010; Resulaj et al., 2009; Yu et al., 2015) and an account of how we sometimes realise that we have made a mistake, even in the absence of external feedback (Charles, King, & Dehaene, 2014; Murphy et al., 2015; van den Berg, Anandalingam, et al., 2016; Yeung & Summerfield, 2012; Yeung, Botvinick, & Cohen, 2004).

Our finding that both pre- and post-decisional evidence impacts confidence and error detection contradicts models that view confidence as reflecting only the balance of evidence up to the point of decision (Kiani et al., 2014; Kiani & Shadlen, 2009; Zylberberg et al., 2012). Similarly, this finding seems difficult to reconcile with *post-decisional locus* models of confidence that make the assumption that confidence reflects only continued processing after the response. In particular, classic standard drift diffusion models (Link, 1975) that assume that a decision is reached when a fixed threshold is crossed, make the prediction that pre-decisional evidence is also constant between trials and therefore that confidence should be determined solely by evidence accumulated post-decisionally.

Although we find that error detection becomes more likely when post-decisional evidence more strongly favours the objectively correct response, as all existing theories would predict (following Rabbitt, 1966a), we find that detection is also more likely following errors that are based on initially weaker evidence. Indeed, our results confirm that the pre-decisional balance of evidence has a lasting impact on decision evaluations made hundreds of milliseconds later. It remains to be established how such results could be reconciled with existing theories that focus solely on post-decisional accumulation of evidence against an initial choice as the core mechanism of error detection (Yeung & Summerfield, 2012; 2014). Our results seem to be more easily accounted for by modified model of first-order decisions (van den Berg, Anandalingam, et al., 2016) which hypothesizes a race between two separate accumulators for each possible decision. According to this view, confidence reflects the

balance of evidence between the competing accumulators, and error detection and changes of mind occur when there is a reversal in the balance of evidence between pre- and post-decisional evidence accumulation, as we observe in our results.

Despite this convergence, it remains to be demonstrated whether this view of confidence as a simple “delayed” first-order decision (Resulaj et al., 2009; van den Berg, Anandalingam, et al., 2016)—i.e., reflecting an evolving balance of evidence that continues to develop even after an initial choice—can entirely account for our findings. Indeed, a recent EEG study suggests that although similar neural signatures of evidence accumulation are apparent before and after the response, post-decisional process differ qualitatively as they accumulate evidence on the likelihood of having made an error rather than votes in favour of one choice or another (Murphy et al., 2015). This evidence converges with theoretical models of confidence as an explicit representation of uncertainty in choice that is distinct from the decision process *per se* (Pouget, Drugowitsch, & Kepecs, 2016), as well as evidence from neuroimaging (Fleming et al., 2010) and neuropsychology (Chua, Pergolizzi, & Weintraub, 2014; Fleming & Lau, 2014) suggesting distinct neural bases for first- and second-order decision processes. The present study does not provide direct evidence for this distinction, but our findings are certainly compatible with the view that confidence does not reflect precisely the same accumulation process as the first-order decision. A valuable extension of the present work would therefore be to investigate how activity in distinct decision- and evaluation-related regions varies with dynamic, stochastic fluctuations in evidence of the kind studied here.

One interesting aspect of our findings is that, perhaps surprisingly, the influence of evidence accumulated both pre- and post-decisionally was apparent regardless of whether instructions emphasized speed or accuracy in responding. Indeed, influence of post-decisional evidence was observed even for correct trials in accuracy blocks, which had the longest

response-times and for which evidence at the time of the response was already high. Although it may appear obvious that an observer should integrate new information to his confidence judgment, an alternative possibility when no time-pressure is applied to the decision could be to wait to reach total certainty before providing a response. It is therefore interesting to observe that this was not the strategy deployed by participants, who appear instead to make an initial choice and to continue to sample the evidence to further evaluate their choice. This result could be considered an artefact of the present experimental design in which post-decisional evidence was always available. However, it could also suggest that integration of evidence preceding and following an initial choice is an essential feature of confidence judgments. As such, this finding seems to contradict the view that integration of post-decisional evidence into confidence judgment occurs only when high speed pressure is applied, forcing participants to produce a response before a decision has truly been reached, as could be suggested by studies that emphasize the role of post-decisional evidence in revising initial decisions (Hilgenstock, Weiss, & Witte, 2014; Moran et al., 2015; Pleskac & Busemeyer, 2010; Yeung & Summerfield, 2012; Yu et al., 2015). More research will be needed to explore whether confidence in itself guides the continuation of information processing after an initial choice is made (Desender, Boldt, & Yeung, 2018) and explore how allowing delayed confidence judgment influences how a first-order decision threshold is set.

At a more detailed level, in both pre- and post-decisional periods, we found an influence on confidence of fluctuations in evidence corresponding to both the chosen and unchosen options. As such, our results seem to conflict with some reports suggesting that confidence, unlike choice, is influenced solely by the strength of evidence favouring the selected option (Koriat, Lichtenstein, & Fischhoff, 1980; Nickerson, 1998; Peters et al., 2017; Zylberberg et al., 2012). Instead, our findings appear to suggest a symmetrical influence on confidence of evidence favouring the two options, consistent with other recent studies (Yu et

al., 2015) and the hypothesis that confidence reflects the balance of evidence between choice options (van den Berg, Anandalingam, et al., 2016). Note that our results do not completely exclude the possibility that participants selectively sampled information to determine their confidence. Indeed, the design of our task allows participants to sample only one of the two boxes to determine which correspond to the high and low mean value. Averaging across trials could then result in an overall effect of symmetry between the selected and the unselected option while participants would in fact sample alternatively one of the stimuli. This interpretation is however unlikely considering the instructions given to the participants to fixate the centre the screen and pay attention to both stimuli. Further research will be needed to explore alternative hypotheses explaining discrepancies between our results and those of earlier studies (Peters et al., 2017). For example, we used a confidence rating scale ranging from Correct to Error, in contrast to a scale from Guess to High confidence in previous studies (Peters et al., 2017; Zylberberg et al., 2012). Perhaps the latter scale leads to a tendency towards confirmation bias in confidence ratings, by not providing participants with the possibility of revising their judgment.

Finally, our results extend previous studies that, like ours, attempt to link error detection and confidence judgments by treating them as part of a single continuum of decision evaluations (Baranski & Petrusic, 1994; Boldt & Yeung, 2015; Scheffers & Coles, 2000). We interpret our findings—of a shared dependence of confidence and error judgments on both pre- and post-decisional evidence—as evidence that they reflect a common underlying metacognitive evaluation process. We favor this interpretation over a possible alternative view, that the shared dependence we observe is an artefact of forcing participants to rate errors and confidence on a single scale, for several reasons. First, by their very definitions, confidence and error judgements fall on a meaningful continuum—a subjective estimate of $p(\text{correct})$ that varies from 0 to 1—rather than being artificially and arbitrarily

forced together (cf. dumping effects in perceptual ratings, e.g., Frank, van der Klaauw, & Schifferstein, 1993). Second, our analyses were not biased to find correlations between pre- and post-decisional evidence and both judgment types—indeed, we did not predict *a priori* that we would see an influence of pre-decisional evidence on error detection, yet our analyses revealed this effect. Finally, previous studies have shown that variations towards both ends of the error-confidence continuum are associated with common neural signatures—graded amplitude changes in well-characterized post-decisional event-related brain potential components (Boldt & Yeung, 2015; Steinhauser & Yeung, 2010).

At a more methodological level, our design shows the distinction between detected and undetected errors in the dynamics of evidence accumulation process: Even in blocks emphasising accuracy, a significant proportion of errors remained undetected while others correctly identified as mistakes, and these trials were associated with differing evidence dynamics as revealed by reverse correlation analysis. This result highlights the importance of allowing confidence judgment to extend beyond “unsure” rating and to allow explicit error detection. Indeed, our pattern of results suggests that classical confidence study which distinguish only “high” and “low” confidence (Fleming & Lau, 2014; Moran et al., 2015; Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010) might miss some important findings by neglecting differences in information processing within the “low confidence” category.

In conclusion, the present study sheds new light on the dynamics of evidence accumulation relating to error-detection and changes of mind, showing that confidence and error judgments integrate information both before and after a decision is produced. These results force us to revise our view on classical models of meta-decision, providing evidence that a common process evaluating the overall signal strength over time can explain error detection, changes of mind and graded confidence judgments.

637 **Acknowledgments**

638 We would like to thank Annika Boldt, Niccolo Pescetelli and Anne-Marike Schiffer for
639 helpful discussion. This project was supported by a postdoctoral grant from the Fondation
640 Fyssen (Paris) to LC. The authors declare no conflict of interest.

References

- Audley, R. J. (1960). A Stochastic Model for Individual Choice Behavior. *Psychological Review*, 67, 1–15.
- Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, 55, 412–428.
- Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 929–945.
- Boldt, A., & Yeung, N. (2015). Shared Neural Markers of Decision Confidence and Error Detection. *Journal of Neuroscience*, 35, 3478–3484.
- Charles, L., King, J.-R., & Dehaene, S. (2014). Decoding the dynamics of action, intention, and error-detection for conscious and subliminal stimuli. *The Journal of Neuroscience*, 34, 1158–1170.
- Chua, E. F., Pergolizzi, D., & Weintraub, R. R. (2014). The Cognitive Neuroscience of Metamemory Monitoring: Understanding Metamemory Processes, Subjective Levels Expressed, and Metacognitive Accuracy Elizabeth. In S. M. Fleming & C. D. Frith (Eds.), (pp. 267–291). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*.
- Desender, K., Boldt, A., & Yeung, N. (2018). Subjective Confidence Predicts Information Seeking in Decision Making. *Psychological Science*, 29, 761–778.
- Festinger, L. (1943). Studies in decision: I. Decision-time, relative frequency of judgment and subjective confidence as related to physical stimulus difference. *Journal of*

664 *Experimental Psychology*, 32, 291–306.

665 Fleming, S. M., & Frith, C. D. (2014). *The cognitive neuroscience of metacognition*. (S. M.
666 Fleming & C. D. Frith, Eds.) *The Cognitive Neuroscience of Metacognition* (Vol.
667 9783642451). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-
668 45190-4

669 Fleming, S. M., & Lau, H. (2014). How to measure metacognition. *Frontiers in Human*
670 *Neuroscience*, 8, 1–9.

671 Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2015). Action-
672 Specific Disruption of Perceptual Confidence. *Psychological Science*, 26, 89–98.

673 Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating Introspective
674 Accuracy to Individual Differences in Brain Structure. *Science*, 329, 1541–1543.

675 Frank, R. A., van der Klaauw, N. J., & Schifferstein, H. N. (1993). Both perceptual and
676 conceptual factors influence taste-odor and taste-taste interactions. *Perception &*
677 *Psychophysics*, 54, 343–54.

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

680 Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy.
681 *Psychological Review*, 18, 186–201.

682 Hilgenstock, R., Weiss, T., & Witte, O. W. (2014). You’d Better Think Twice: Post-Decision
683 Perceptual Confidence. *NeuroImage*, 99, 323–331.

684 Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence
685 in humans and animals. *Philosophical Transactions of the Royal Society B: Biological*
686 *Sciences*, 367, 1322–1337.

687 Kepecs, A., & Mainen, Z. F. (2014). A Computational Framework for the Study of
688 Confidence Across Species. In S. M. Fleming & C. D. Frith (Eds.), *The Cognitive*
689 *Neuroscience of Metacognition* (pp. 115–145). Berlin, Heidelberg: Springer Berlin
690 Heidelberg.

691 Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates,
692 computation and behavioural impact of decision confidence. *Nature*, *455*, 227–231.

693 Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both
694 evidence and decision time. *Neuron*, *84*, 1329–1342.

695 Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision
696 by neurons in the parietal cortex. *Science*, *324*, 759–764.

697 Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of*
698 *Experimental Psychology: Human Learning & Memory*, *6*, 107–118.

699 Link, S. (1975). The relative judgment theory of two choice response time. *Journal of*
700 *Mathematical Psychology*.

701 Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating
702 metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*,
703 422–430.

704 Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a
705 causal determinant of confidence: Novel data and a computational account. *Cognitive*
706 *Psychology*, *78*, 99–147.

707 Murphy, P. R., Robertson, I. H., Harty, S., & O’Connell, R. G. (2015). Neural evidence
708 accumulation persists after choice to inform metacognitive judgments. *eLife*, *4*, 1–23.

709 Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises.

710 *Review of General Psychology*, 2, 175–220.

711 Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., ... Lau,
 712 H. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain.
 713 *Nature Human Behaviour*, 1, 1–8.

714 Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of
 715 choice, decision time, and confidence. *Psychological Review*, 117, 864–901.

716 Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct
 717 probabilistic quantities for different goals. *Nature Neuroscience*, 19, 366–374.

718 Rabbitt, P. M. (1966). Error correction time without external error signals [55]. *Nature*.

719 Rabbitt, P. M. (1966). Errors and error correction in choice-response tasks. *Journal of*
 720 *Experimental Psychology*, 71, 264–272.

721 Rabbitt, P. M., & Vyas, S. (1981). Processing a display even after you make a response to it.
 722 how perceptual errors can be corrected. *Quarterly Journal of Experimental Psychology*,
 723 33, 223–239.

724 Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in
 725 decision-making. *Nature*, 461, 263–266.

726 Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., & Lau, H. (2010). Theta-burst
 727 transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual
 728 awareness. *Cognitive Neuroscience*, 1, 165–175.

729 Scheffers, M. K., & Coles, M. G. H. (2000). Performance monitoring in a confusing world:
 730 Error-related brain activity, judgments of response accuracy, and types of errors. *Journal*
 731 *of Experimental Psychology: Human Perception and Performance*, 26, 141–151.

732 Steinhauser, M., & Yeung, N. (2010). Decision Processes in Human Performance

733 Monitoring. *Journal of Neuroscience*, 30, 15643–15653.

734 van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert,
735 D. M. (2016). A common mechanism underlies changes of mind about decisions and
736 confidence. *eLife*, 5, 1–21.

737 van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016).
738 Confidence Is the Bridge between Multi-stage Decisions. *Current Biology*, 26, 3157–
739 3168.

740 Vickers, D. (2001). Where does the balance of evidence lie with respect to confidence.
741 *Proceedings of the Seventeenth Annual Meeting of the International Society for*
742 *Psychophysics*, 148–153.

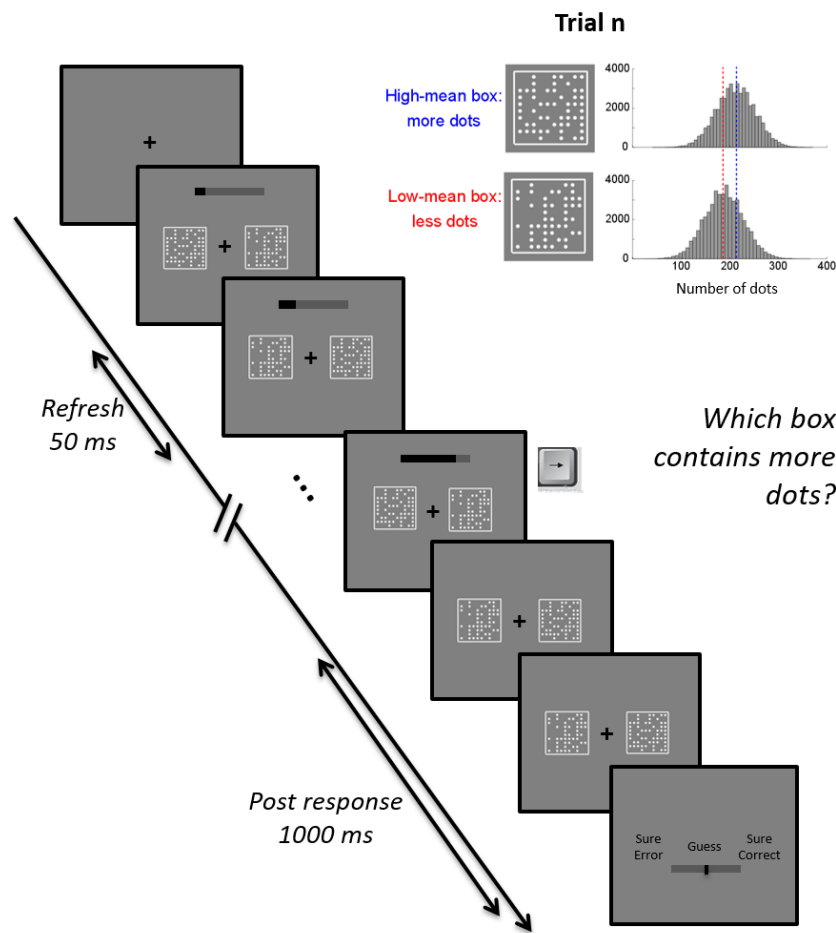
743 Vickers, D., & Packer, J. (1982). Effects of alternating set for speed or accuracy on response
744 time, accuracy and confidence in a unidimensional discrimination task. *Acta*
745 *Psychologica*, 50, 179–197.

746 Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection:
747 Conflict monitoring and the error-related negativity. *Psychological Review*, 111, 931–
748 959.

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

752 Yu, S., Pleskac, T. J., & Zeigenfuse, M. D. (2015). Dynamics of postdecisional processing of
753 confidence. *Journal of Experimental Psychology: General*, 144, 489–510.

754 Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a
755 perceptual decision. *Frontiers in Integrative Neuroscience*, 6, 1–10.



757

758 *Figure 1: Experimental procedure. Stimuli consisted of a stream of images updated every 50*
 759 *ms displaying two boxes containing dots at random positions in a 20 x 20 array. The*
 760 *participants' task was to determine which of the two boxes contained more dots on average.*
 761 *The number of dots in each box was drawn from two Normal distributions centred on a high*
 762 *(212 dots) and a low value (188 dots). The time the participant had left to respond was*
 763 *indicated by a bar on the top of the screen that gradually filled up. Participants were*
 764 *instructed whether the bar would fill slowly ("Accuracy Block") or quickly ("Speed Block")*
 765 *at the beginning of each block. Importantly, the stimulus stream continued to be displayed for*
 766 *1000 ms after each response. Participants were then asked to rate the confidence they had in*

767 *their response on a scale going from “Sure I made an Error” to “Sure I responded*
768 *correctly”.*
769

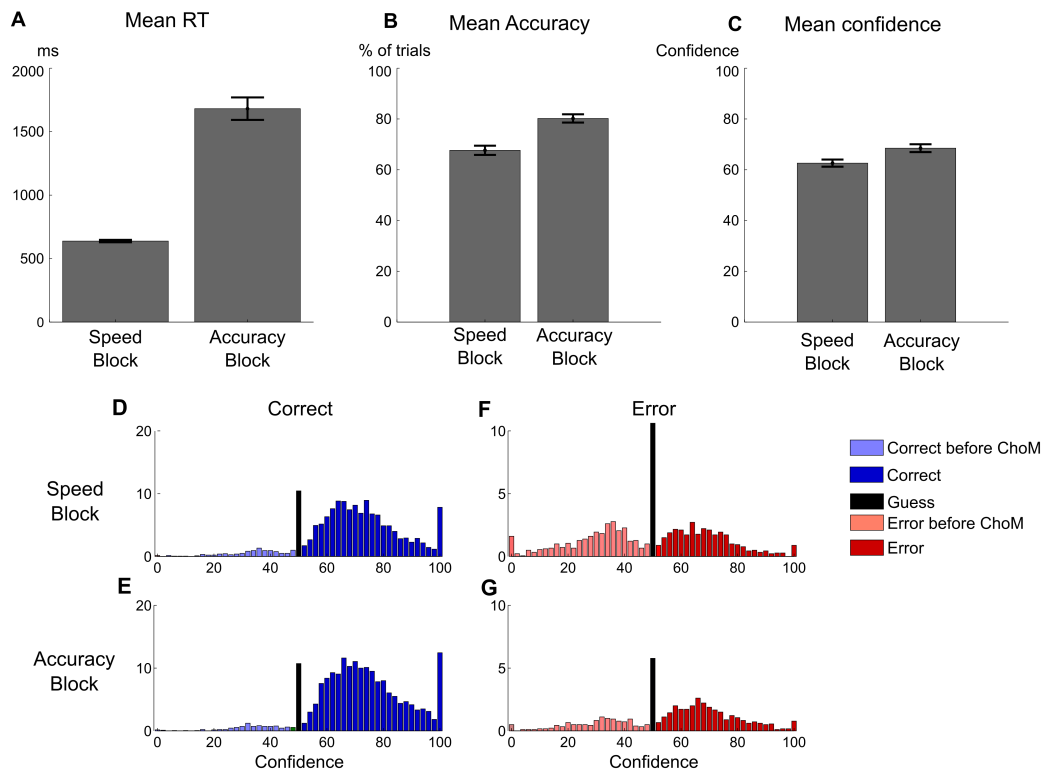
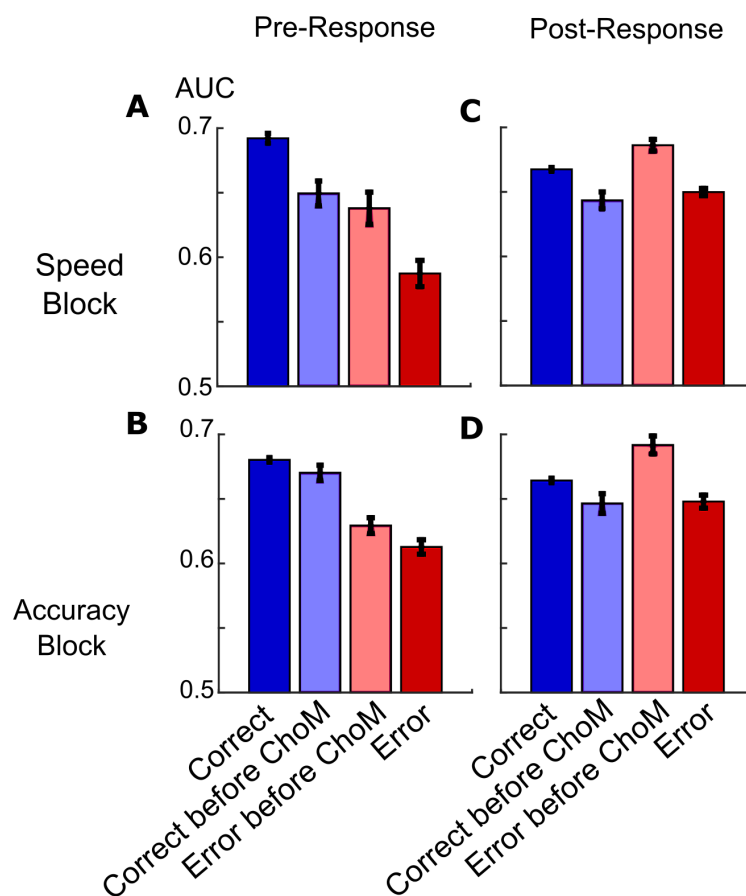


Figure 2: Response Time, accuracy and confidence for both types of blocks. A-C: Mean response-time, accuracy and confidence for Speed and Accuracy Blocks. D-G: Average distribution over participants of the use of the confidence scale for Correct (D-E) and Error trials (F-G) separately for Speed (D-F) and Accuracy blocks (E-G), with proportion of certain correct (confidence = 100%), perceived correct (50% < confidence < 100%), guess (confidence = 50%), changes of minds (ChoM: 0% < confidence < 50%), and certain error (confidence = 0%) trials separately for Correct and Error trials.

781



782

783

784 *Figure 3: Quantifying objective levels of evidence in pre- and post-response time-intervals.*

785 *For each trial, AUC values computed from the ROC curve associated with the distribution of*

786 *dot numbers in the low-mean and high-mean boxes were computed for the pre-response (A-*

787 *B) and the post-response (C-D) time interval, separately for Speed (A,C) and Accuracy (B,D)*

788 *blocks. The obtained values were averaged according to accuracy and metacognitive*

789 *accuracy: Correct trials perceived as correct (dark blue), Correct with later change of mind*

790 *(ChoM, light blue), Errors followed by a change of mind (light red) and Errors that remained*

791 *undetected (dark red).*

792

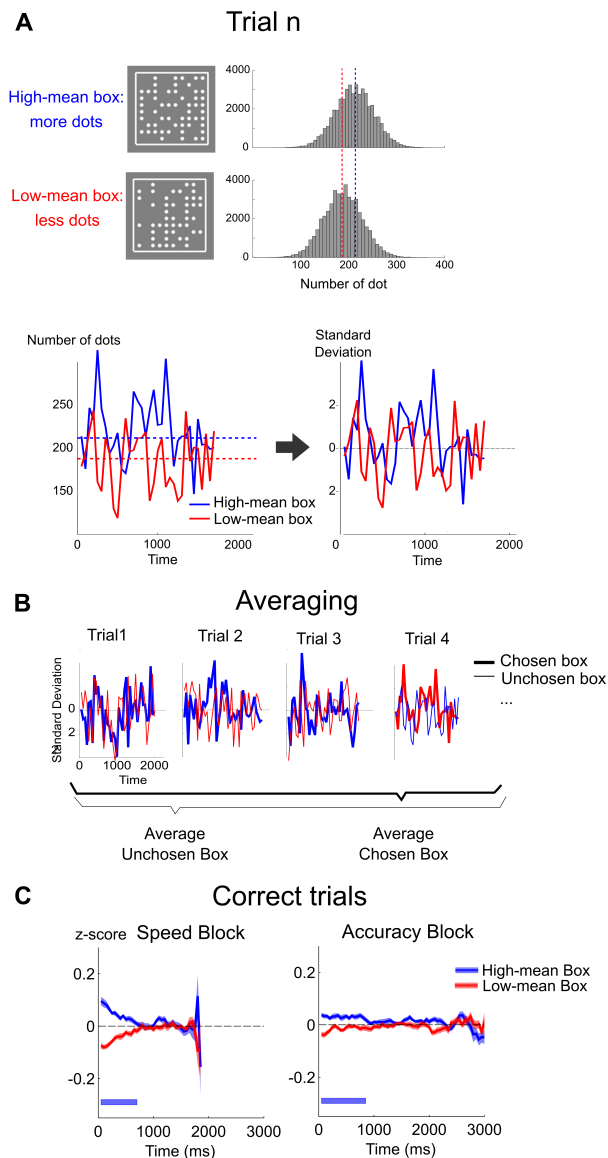


Figure 4: Reverse correlation method. (A) An example of the stimuli on one sample and the distribution across time-samples on a trial n of the number of dots in each type of stimulus (increased for display purposes). Bottom panel shows an example time-course of the number of dots in the low-mean and high-mean boxes, before and after normalization. (B) Reverse correlation results are then obtained by averaging together across trials the time-courses, according to which stimulus was chosen by the participant on each trial. (C) Example of average reverse correlation results obtained for Correct trials detected as correct, for Speed and Accuracy blocks.

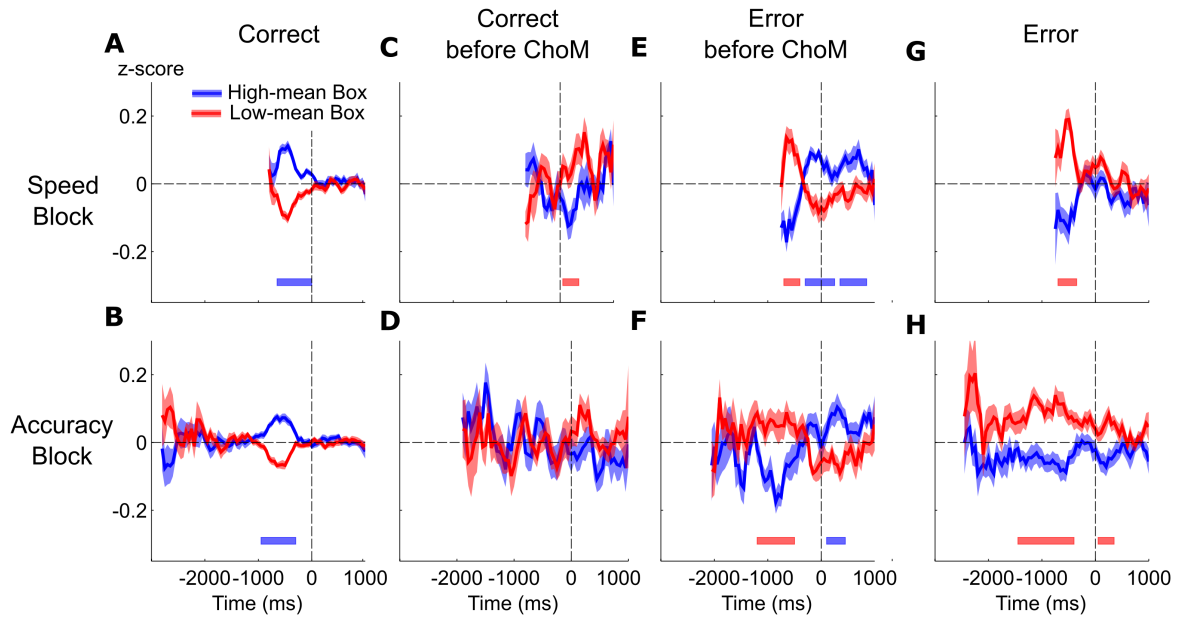


Figure 5: Influence of sensory evidence on response and error-detection. Each panel depicts evidence fluctuations time-locked to the response (vertical dotted line at 0 ms), showing the normalized number of dots in the low-mean box (red line) and high-mean box (blue line), in Speed (A,C,E,G) and Accuracy blocks (B,D,F,H) for Correct trials correctly classified as Correct (A-B), Correct responses followed by change of mind (ChoM, C-D), Errors followed by change of mind (E-F) and Error without a change of mind (G-H). Significance across participants of the difference between the two curves is indicated by colored lines at the bottom of the graph, with blue lines indicating a positive difference between the correct and the incorrect stimulus and with red lines indicating a negative difference.

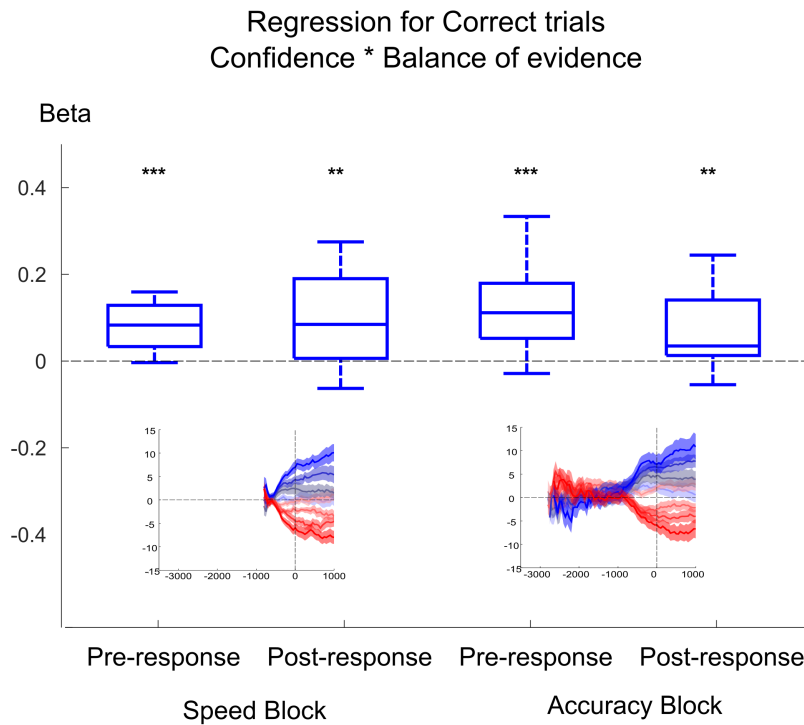


Figure 6: Correlation between level of evidence and confidence in correct trials. The figure depicts the boxplot (central line: median, bottom and top lines: 25th and 75th quantiles, whiskers: most extreme data points not considered as outliers) of the betas of individual regression across trials between the average balance of evidence between the two stimuli in the pre- and post-decisional time interval for Speed and Accuracy blocks. Insert depicts the cumulative sum of the evidence in the low-mean (red) and the high-mean (blue) box according to confidence bin (darker color = higher confidence). Stars indicate significant difference from zero with $p < 0.05$:*, $p < 0.001$:**, $p < 0.0001$:***.

Supplementary material

Dynamic evidence accumulation supporting confidence judgments and error detection

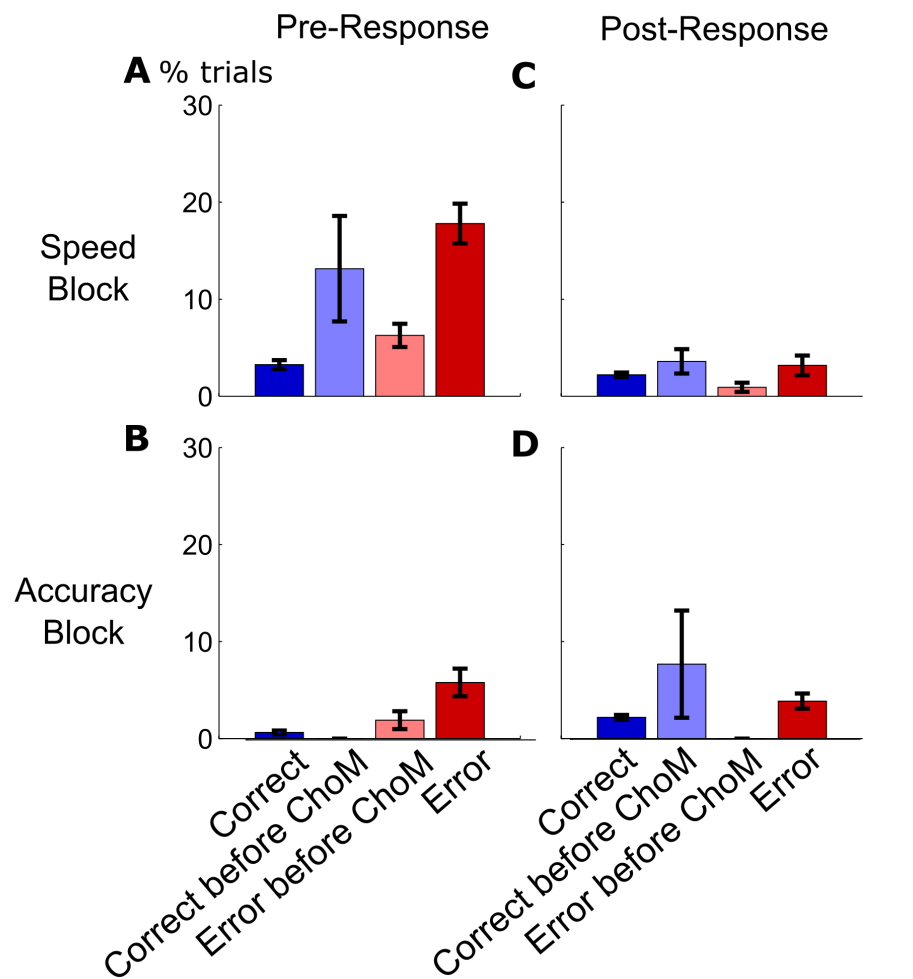
Lucie CHARLES , Nick YEUNG

Methods

Non-parametric statistics for reverse correlation analysis

For each time-sample, non-parametric Mann–Whitney U tests were performed testing the difference in the normalized number of dots between the two boxes. Time clusters were then identified by taking all dyads of time-samples adjacent in time with $p < 0.05$. Final significance of each time cluster was determined by computing the sum of the z-score values of the entire cluster, and comparing it with the results of Monte-Carlo permutations (2000 permutations). Clusters were considered significant at corrected $p < 0.05$ if the probability computed with the Monte-Carlo method was less than 5% (one-tailed test). Significant clusters at each time-point are shown color coded at the bottom of each relevant graph (Figure 5), with blue indicating a significant evidence in favor of the high-mean over the low-mean box and red indicating the reverse.

19 Supplementary Figure



20

21 *Figure Suppl: Proportion of trials in which evidence favored the incorrect response during the*
 22 *pre-response (A-B) and the post-response (C-D) time interval, for Fast (A,C) and Slow (B,D)*
 23 *blocks. The obtained values were averaged according to accuracy and metacognitive accuracy:*
 24 *Correct trials perceived as correct (dark blue), Correct with later change of mind (dark purple),*
 25 *Errors followed by a change of mind (lighter purple) and Errors that remained undetected*
 26 *(bright red).*