Manipulating prior beliefs causally induces under- and overconfidence

Hélène Van Marcke*[1], Pierre Le Denmat*[1], Tom Verguts[2], & Kobe Desender[1]

1. Brain and Cognition, KU Leuven, Belgium; 2. Department of Experimental Psychology, Ghent University, Belgium

* These authors contributed equally

Correspondence: Helene.VanMarcke@kuleuven.be

Abstract

Making a decision is invariably accompanied by a sense of confidence in that decision. There is widespread variability in the exact level of confidence, even for tasks that do not differ in objective difficulty. Such expressions of under- and overconfidence are of vital importance, as they are related to fundamental life outcomes. Yet, a clear account regarding the computational mechanisms underlying under- and overconfidence is currently missing. In the current work, we propose that prior beliefs in the ability to perform a task can explain why confidence can differ dramatically despite similar task performance. In two experiments, we provide evidence for this hypothesis by showing that manipulating prior beliefs about task performance in an induction phase causally influences reported levels of confidence in a testing phase, while leaving actual performance unaffected. This is true both when prior beliefs are influenced via manipulated feedback and by manipulating the difficulty of the task during the training phase. These results are accounted for within an accumulation-to-bound model by assuming an additional parameter controlling prior beliefs about task performance. Our results provide a fundamental mechanistic insight into the computations underlying over- and underconfidence.

Introduction

Human decision making is accompanied by a sense of confidence regarding the accuracy of the chosen options. In experimental work, this sense of confidence usually correlates with objective accuracy, with participants reporting high confidence for correct choices and low confidence for incorrect choices (Fleming et al., 2010). This tight link between confidence and accuracy is captured in the theoretical proposition that confidence reflects the probability of a choice being correct (Sanders et al. 2016; Kiani & Shadlen, 2009). An assumption derived from these theoretical views is that humans should be rather stable in their way of computing and reporting confidence. For example, several studies have shown that confidence can be captured by a model that does not feature additional free parameters to capture potential biases in confidence (Sanders et al. 2016; Schmack et al., 2021). However, while decision confidence is on average well explained by models quantifying confidence as the probability of a choice being correct, there is clear evidence of vast differences between individuals concerning the actual reported level of confidence (Ais et al., 2015). This is clearly visible in simple, low-level perceptual decision-making tasks, where some subjects systematically underestimate and others overestimate their choice accuracy, an observation often referred to as under- and overconfidence, respectively. Although expressions of over- or underconfidence may not have important consequences in the context of a laboratory task, such biases can have far-reaching implications in real life. For example, overconfidence has been related to increased sharing of fake news (Lyons et al., 2021), with diagnostic inaccuracies in physicians (Saposnik et al., 2016), and more generally with low self-esteem (Moses-Payne 2019). Moreover, impaired insight into the accuracy of one's decisions (or, impaired metacognition) has been linked to holding radical beliefs (Rollwage, 2018) and a variety of psychiatric symptoms (for a review, see Hoven et al., 2019).

Although there is clear evidence for individual differences in confidence, with potentially far-reaching consequences, the origin of these differences is ill understood. Although some researchers have proposed explanations in terms of impression management (Schwardman & Van der Weele, 2020; Belmi et al., 2020) or exposure to feedback (Baranski & Petrusic, 1994; Petruzello & Corbin, 1988), these accounts do not offer a fundamental understanding of the underlying mechanisms. Therefore, in the current work we investigated whether it is possible to account for biases in confidence within probabilistic frameworks. Simply put: If confidence indeed reflects the probability of being correct, then why do different individuals report such vastly different estimates when asked to evaluate the accuracy of their own choices? To explain the computational dynamics behind over- and underconfidence, we leveraged an often under-appreciated aspect about probabilistic models of confidence, namely that the probability of being correct itself depends on prior beliefs. Everything else being equal, the probability of making a correct choice is larger in an easy versus a difficult task. Intuitively: If you are performing an easy task, the probability of being correct on a specific trial is higher than when you are performing a hard task, simply because in an easy task there are more correct trials (see **Figure 1**). Although we are not the first to acknowledge the role of priors in the computation of confidence (Moreno-Bote, 2010; Drugowitsch et al., 2014; Khalvati et al., 2021), so far empirical support for this notion has been lacking. Therefore, in the current work, we aimed to provide direct evidence that prior beliefs underlie under- and

overconfidence by explicitly manipulating prior beliefs about task performance. In two experiments, we manipulated prior beliefs during the training phases and looked at the influence of these manipulations on confidence ratings during a subsequent testing phase. Our results showed that an alteration of prior beliefs, either by means of social feedback or by means of training on tasks of different levels of difficulty, selectively affected subsequent confidence ratings, while leaving performance unaffected. Importantly, these effects of prior beliefs on decision confidence were easily accounted for by a model incorporating prior belief.

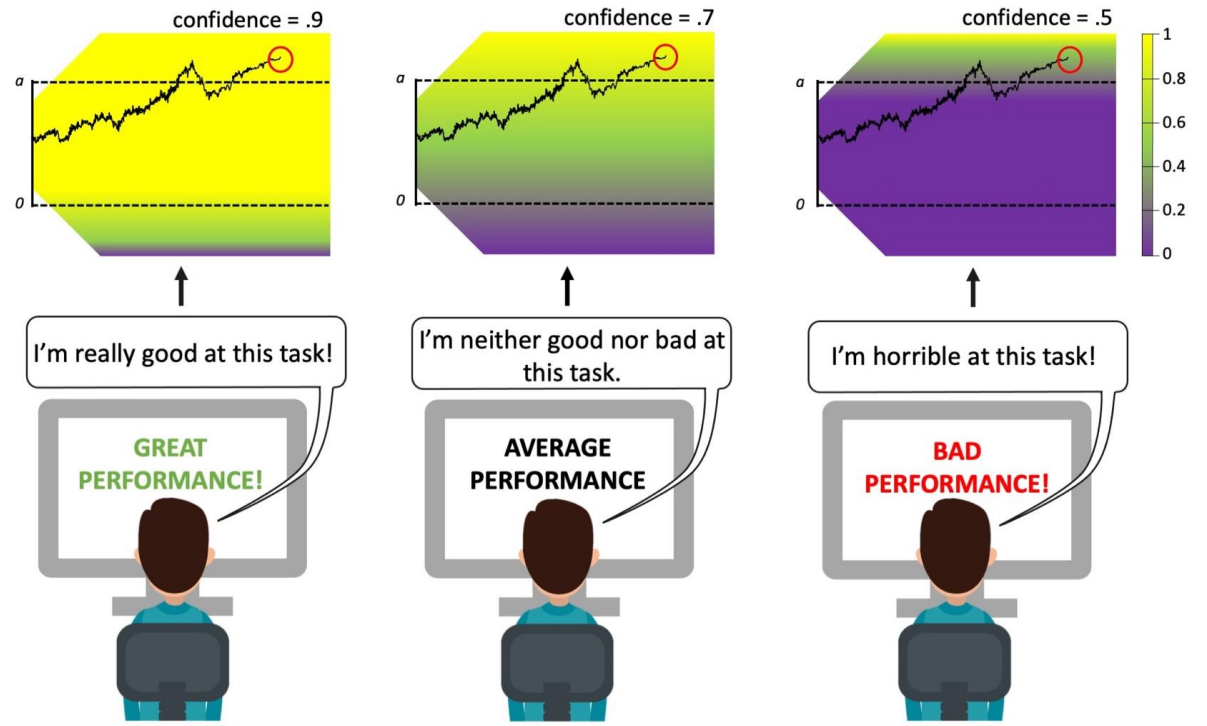


Figure 1. Cartoon figure illustrating how prior beliefs might influence decision confidence*. We hypothesised that participants exposed to feedback indicating that they are doing great vs. performing very poorly, will create the beliefs that they are good vs. bad at a task. Critically, this will influence their prior beliefs about the probability of making correct responses in this task. As a consequence, the exact same trial can result in different levels of confidence (i.e. reflecting under- and overconfidence) .*

Results

To unravel the influence of prior beliefs on decision confidence, we carried out two experiments that aimed to causally influence participants' prior beliefs about their ability to accurately perform the task. In both experiments, participants performed three similar perceptual decision making tasks. Each task started with a training phase where we manipulated participants' prior beliefs in their ability to accurately perform the task. This was done by providing them with feedback indicating that their performance was good, average or poor. In the subsequent test phase of each task, we tested the influence of the manipulation on trial-by-trial confidence ratings. We hypothesised that this manipulation of prior beliefs would influence confidence ratings during the following test phase, indicating a causal influence of prior beliefs on confidence ratings.

Experiment 1: Manipulating prior beliefs via comparative feedback causally induces under- and overconfidence. In Experiment 1 (N=50), we used comparative feedback to influence prior beliefs about task performance. Participants were told that they would receive feedback every 24 trials about their performance on the task, relative to a group of participants who performed the same task at an earlier time. Unbeknown to participants, feedback was manipulated so that for one task it indicated that performance was better than most participants, that it was on average for another task, and worse than most participants for the third task (see **Figure 2**). Afterwards, participants took part in a test phase during which they no longer received feedback but instead rated their perceived level of confidence on each trial. In line with our main hypothesis, confidence ratings during the test phase depended on the feedback that participants received during the training phase, $F(2,47) = 16.65$, $p < .001$ (see **Figure 3A**). Participants reported a higher level of confidence in their choices after being exposed to feedback saying they performed better (M = 4.79), average (M = 4.64) or worse (M = 4.41) compared to the reference group. Apart from this, there was the expected effect of trial difficulty on confidence ratings, $F(2,47) = 159.71$, $p < .001$. There was also a small interaction between feedback condition and trial difficulty $F(4,30744) = 2.60$, $p = .034$, indicating a differential effect of feedback on confidence according to the difficulty level. However, as can be seen on **Figure 2**, our main effect was still notably visible for all difficulties. Importantly, our induction of prior beliefs selectively affected decision confidence, whereas it left objective performance unaffected. Both accuracy and reaction times were affected by trial difficulty (accuracy: $\chi^2(2) = 2438.88$, $p < .001$; RTs: $F(2,31430) = 315.03$, $p < .001$), but did not vary by feedback condition (accuracy: $\chi^2(2) = 1.4$, $p = .05$; RTs: $F(2,48) = 1.97$, $p = .15$; see **Figure 3A**).

Interestingly, the effect of prior beliefs on confidence was quite persistent throughout the testing phase. Each testing phase comprised three blocks of 72 trials, separated by a break of one minute (see Methods). When adding the factor "block" to the model discussed above, there was a significant interaction between block and feedback condition, $F(4,31412) = 4.98$, $p < .001$, showing a subtle decrease in the effect across time. However, the effect was still remarkably consistent in each of the three blocks (block 1: $F(2,48) = 21.79$, $p < .001$; block 2: $F(2,48) = 14.20$, $p < .001$; block 3: $F(2,48) = 9.51$, $p < .001$).

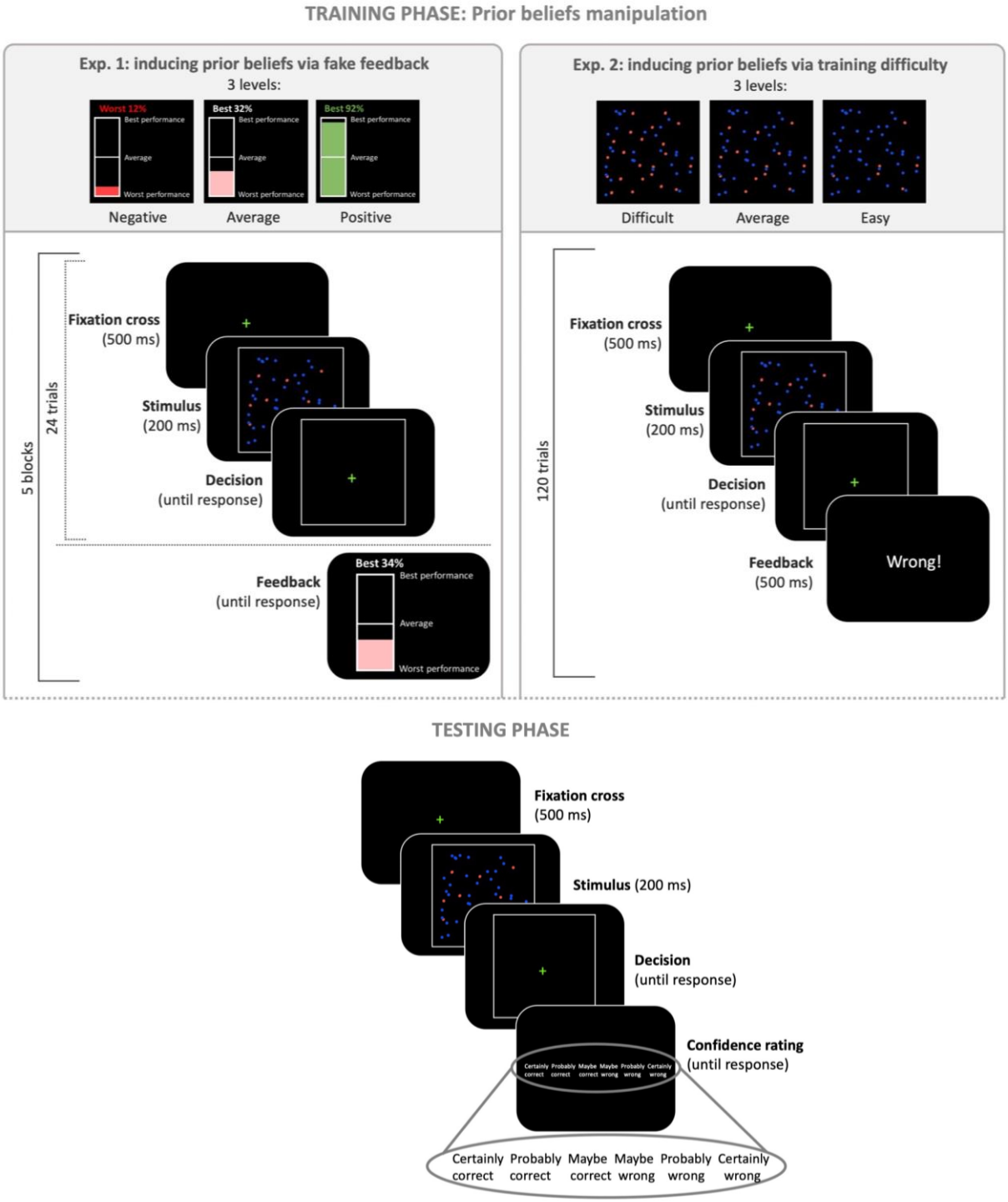


Figure 2. Experimental design. *In both experiments, participants performed three different perceptual decision-making tasks. Each task started with a training phase during which prior beliefs were induced via manipulated feedback (Experiment 1) or a different level of task difficulty (Experiment 2). After the training phase, participants completed a test phase during which they received no feedback but instead rated their decision confidence on each trial.*

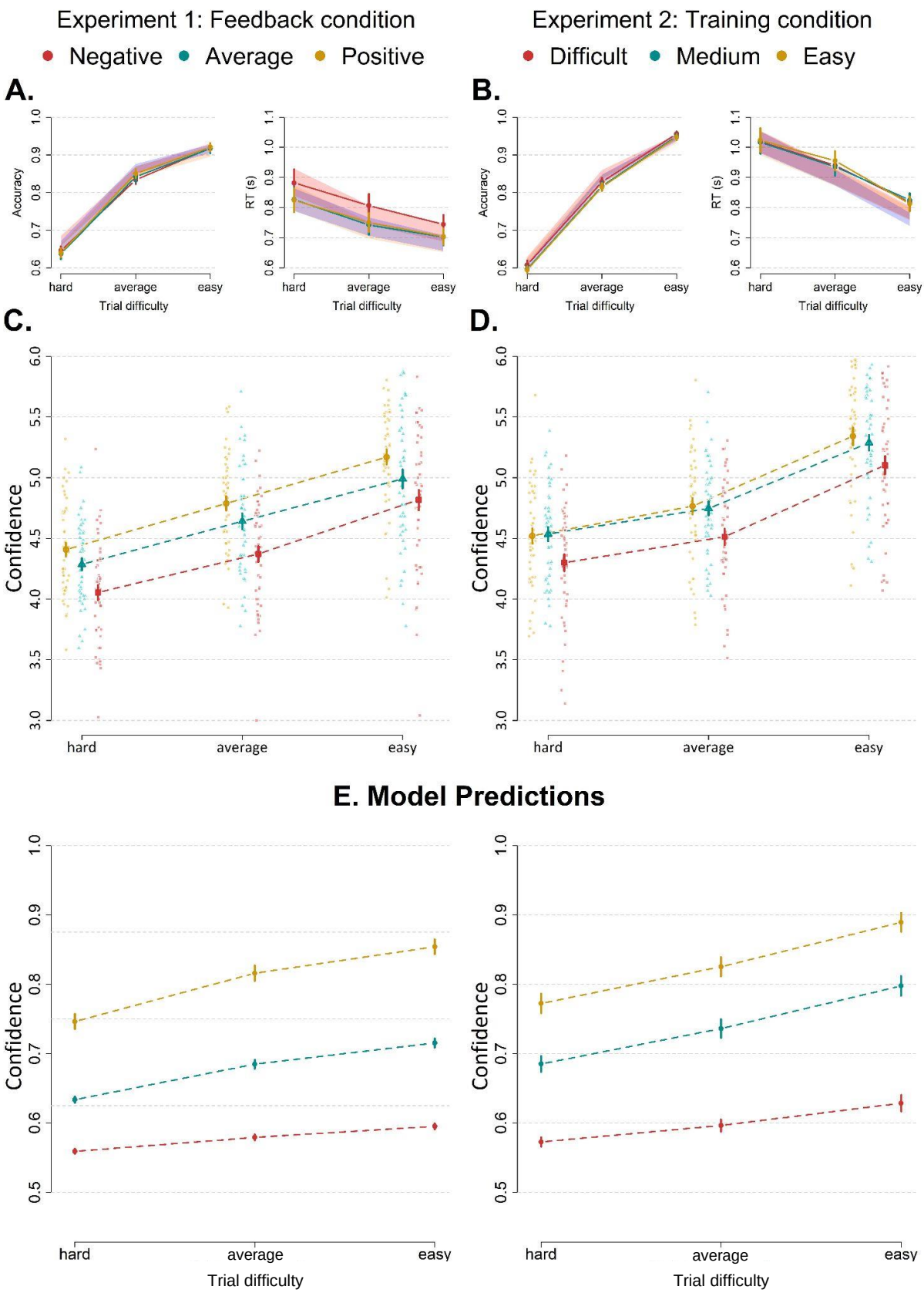


Figure 3. Manipulating prior beliefs causally induces under- and overconfidence. A. *In Experiment 1 (left column), providing participants with feedback that they were performing better, equal or worse than a reference group left objective performance unaffected (top row), but induced over- and underconfidence, respectively*

(middle row). This induction effect on confidence was captured by our computational model (bottom row). ***B.*** *These results were replicated in Experiment 2 (right column), where training difficulty was manipulated. Note: shaded bars in A-B reflect model fits, error bars reflect standard error of the mean, small dots in C-D reflect individual participants.*

Experiment 2: Manipulating prior beliefs via differences in training difficulty. In Experiment 2 (N=50), we altered prior beliefs about task performance ability by varying the difficulty of the task during the training phase. Participants were only trained on easy trials on one task, on trials of medium difficulty on another task and on difficult trials on a third task (Figure 2B). Contrary to Experiment 1, participants received genuine feedback about their choice accuracy ("wrong" or "false") on every trial. Critically, because the tasks differed in difficulty we achieved a similar feedback pattern to that of Experiment 1: On average, participants saw positive feedback on 95% of the trials (near perfect performance), on 79% of the trials (average performance), or on 60% of the trials (slightly above chance performance) for the three tasks, respectively. Despite the different approach in altering prior beliefs, the results fully replicated those of Experiment 1: Confidence ratings of the testing phase depended on the trial difficulty of the training phase, $F(2,46) = 8.19$, $p < .001$ (see **Figure 3B**). Participants reported higher levels of confidence after being previously trained on an easy task ($M = 4.88$), versus a task of medium difficulty level ($M = 4.86$) versus a difficult task ($M = 4.64$). As expected, trial difficulty during the testing phase had an effect on confidence ratings, $F(2,30109) = 2122.11$, $p < .001$, with no interaction between both, $F(4,30109) = 1.64$, $p = .16$. Again, our manipulation left task performance unaffected. Accuracy and RTs were significantly influenced by trial difficulty (accuracy: $\chi^2(2) = 3089.45$, $p < .001$; RTs: $F(2, 30109) = 563.52$, $p < .001$), but not by training condition (accuracy: $\chi^2(2) = 2.35$, $p = .31$; RTs: $F(2,46) = 0.01$, $p = 0.99$), nor was there an interaction (accuracy: $\chi^2(2) = 1.6$, $p = .81$; RTs: $F(2, 30109) = 1.52$, $p = .19$; see **Figure 3B**).

Similar to Experiment 1, the influence of prior beliefs on confidence was persistent across time. When adding block to the analysis on confidence reported earlier, there was no significant interaction between training condition and block, $F(4,30091) = 2.3$, $p = 0.056$, and the effect was remarkably consistent across the three blocks (block 1: $F(2,80) = 15.97$, $p < .001$; block 2: $F(2,81) = 9.83$, $p < .001$; block 3: $F(2,82) = 10.69$, $p < .001$).

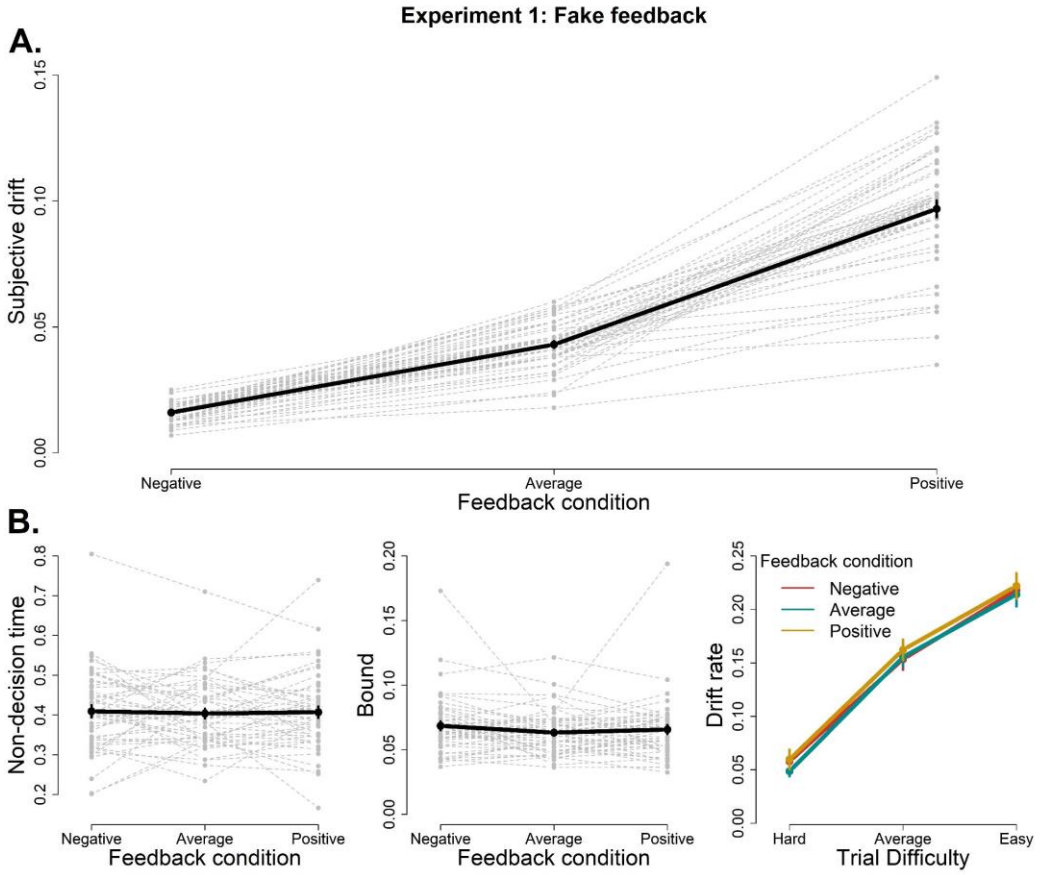
Introducing prior beliefs into models of decision confidence. In order to model the underlying mechanisms by which prior beliefs influence the reported level of confidence, we turned towards computational models of decision confidence. We focused on accumulation-to-bound models, a family of models that have successfully accounted for choices, reaction time and confidence (Kiani & Shadlen, 2009; Zylberg et al., 2016). The main tenet of accumulation-to-bound models is that choices are described as the noisy accumulation of evidence that terminates once one of two opposite decision boundaries is hit, at which point a response is triggered. The rate of evidence accumulation is controlled by the drift rate parameter (v), representing the efficiency of information extraction from the stimulus. In order to account for decision confidence within such a model, it has been argued that confidence reflects the probability of a choice being correct, conditional on the the state of the accumulator (i.e. the amount of evidence accumulated), the decision time and the choice (Kiani et al., 2014; Moreno-Bote, 2010; Desender et al, 2021). In **Figure 1**, this is represented by the heatmaps that

visualise how different combinations of evidence (y-axis) and time (x-axis) are associated with different levels of confidence (darker colours are associated with higher confidence). This representation of probability correct can be thought of as the participants' prior belief about their performance at a task. When a participant believes that they are very bad at a task, this can be thought of as reducing the probability of being correct in this 2D space. In previous work using this approach, it is often (implicitly) assumed that participants have learned the association between evidence, reaction time and probability of being correct via experience, and therefore their reported probability of being correct corresponds to the actual average choice accuracy. Put differently, in previous work it was assumed that prior beliefs about task performance correspond to actual task performance. In practice, however, individuals only have an imperfect approximation of the probability of being correct. For example, participants in Experiment 1 who receive mostly positive or mostly negative feedback will come to believe that their probability of being correct is rather high or low, respectively. As such, manipulated prior beliefs can influence the computed level of decision confidence.

Rather than assuming a correspondence between beliefs about performance and actual performance, we added an additional parameter reflecting the prior belief about the ability to perform a task correctly. This prior belief parameter allows us to distinguish how individuals *think* they perform from how individuals *actually* perform. Formally, whereas task performance in evidence accumulation models is controlled by the drift rate, we introduce a prior beliefs parameter that can be thought of as a *subjective* drift rate, reflecting how good participants *think* they are at a task (see Methods for full details). There is a direct relation between this prior beliefs parameter and the computation of the probability of being correct, which explains how conditions with identical task performance (i.e., same drift rates) but different prior beliefs (i.e. different subjective drift rates) can lead to differences in subjective confidence.

Modelling the effect of prior beliefs on decision confidence. In order to demonstrate that our accumulation-to-bound model can explain under- and overconfidence as the results of differences in prior beliefs, we queried predictions from a model exposed to the same feedback as participants during the induction phase. We estimated which prior belief parameter provided confidence predictions during the training that best matched the feedback that participants received. As expected, when the model was exposed to negative feedback, the estimated prior belief parameter was lower compared to when it was exposed to positive feedback, $F(2,94) = 450.1$, $p < .001$ (Experiment 1, **Figure 4A**). Likewise, when the model was trained on an easy task, the estimated prior belief parameter was lower compared to when the model was trained on a difficult task, $F(2,92) = 38.04$, $p < .001$ (Experiment 2, **Figure 4C**). The other estimated parameters of our accumulation-to-bound model fitted to the testing phase data (drift rate, bound, non-decision time) did not vary with the feedback conditions in Experiment 1 (all *p*s > .36, **Figure 4B**), nor were they influenced by the training difficulty manipulation in Experiment 2 (all *p*s > .31, **Figure 4D**). Thus, our model successfully captured the influence of manipulated feedback (Experiment 1) and differences in task difficulty (Experiment 2) on the parameter capturing the prior belief about task performance.

Next, we tested whether the difference in prior belief induced during the induction phase was sufficient to capture under- and overconfidence in the testing phase. To do so, we queried predictions from our model based on the accumulation-to-bound parameters of the test phase, but using the prior belief created during the induction phase. As expected, for both Experiment 1 and Experiment 2 (**Figure 3E**) the model predicted an increase of confidence with lower trial difficulty, (Experiment 1: $F(2,282) = 344.91$, $p < .001$; Experiment 2: $F(2,276) = 472.20$, $p < .001$), and more importantly increases of confidence with increasingly positive feedback ($F(2,47) = 274.48$, $p < .001$, Experiment 1), and lower task difficulty during the training phase ($F(2,46) = 98.54$, $p < .001$, Experiment 2). Contrary to the empirical data, there was an interaction between both factors, $p < .001$. In sum, we successfully accounted for expressions of under- and overconfidence within accumulation-to-bound models by taking into account prior beliefs.


Experiment 1: Fake feedback
A.
Subjective drift
0.00
0.05
0.10
0.15
Negative
Average
Positive
Feedback condition
B.
Non-decision time
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Negative
Average
Positive
Feedback condition
Bound
0.00
0.05
0.10
0.15
0.20
Negative
Average
Positive
Feedback condition
Drift rate
0.00
0.05
0.10
0.15
0.20
0.25
Feedback condition
Negative
Average
Positive
Hard
Average
Easy
Trial Difficulty

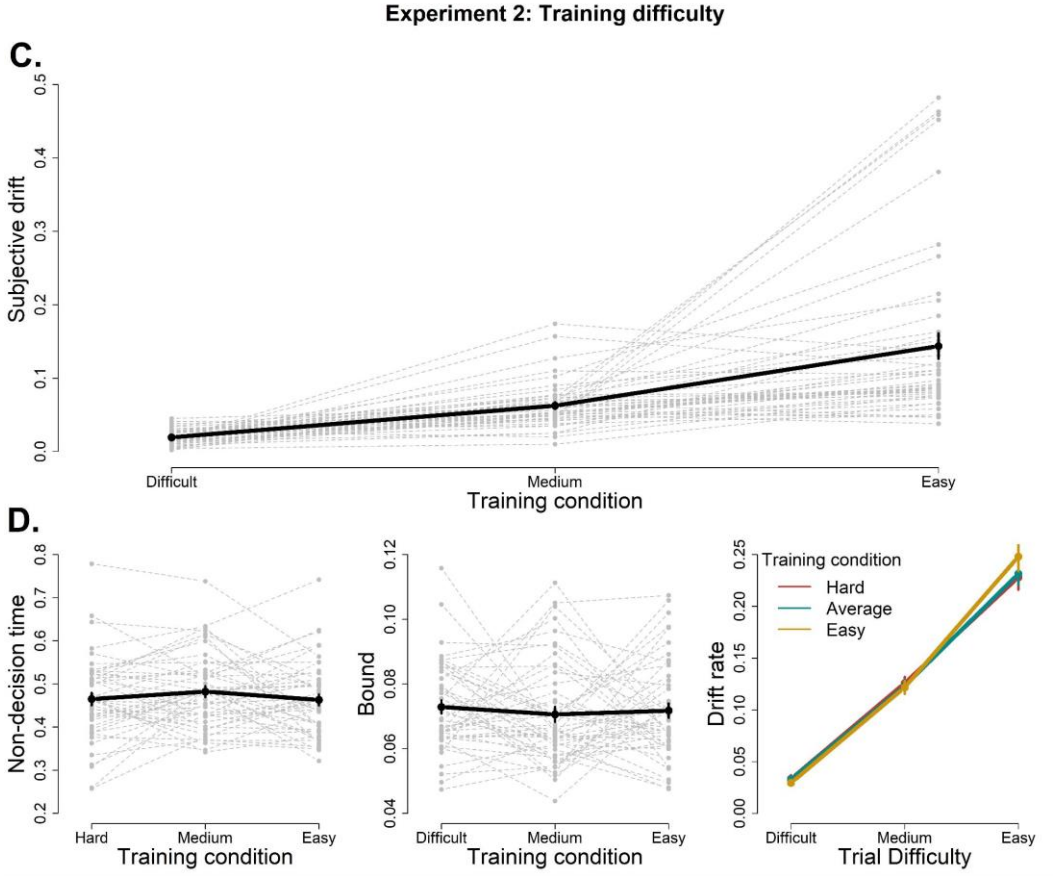

Experiment 2: Training difficulty
C.
Subjective drift
0.0
0.1
0.2
0.3
0.4
0.5
Difficult
Medium
Easy
Training condition
D.
Non-decision time
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Hard
Medium
Easy
Training condition
Bound
0.04
0.06
0.08
0.10
0.12
Difficult
Medium
Easy
Training condition
Drift rate
0.00
0.05
0.10
0.15
0.20
0.25
Training condition
Hard
Average
Easy
Difficult
Medium
Easy
Trial Difficulty

Figure 4. The prior beliefs manipulations were captured by our prior belief parameter. A. *In Experiment 1, our prior belief parameter increased when participants were provided positive feedback.* ***B.*** *The feedback conditions in the training phase did not influence the other parameters of the computational model.* ***C-D.*** *Similarly, in Experiment 2, only the prior belief parameter was sensitive to the training difficulty. Note: grey dotted lines reflect individual participants, solid black lines represent the mean, error bars reflect standard error of the mean.*

Discussion

Humans differ greatly in the degree to which they are confident in their choices. These differences in the overall level of confidence, often referred to as under- and overconfidence, are of vital importance as they show marked relations with clinical symptomatology such as anxiety and depression. Here, we provide direct evidence that these differences arise from prior beliefs about the ability to perform a task. In two experiments, we have shown that a manipulation of prior beliefs causally influences the reported level of decision confidence. Furthermore, we have shown that these findings are easily accounted for within a dynamic probabilistic model of decision confidence, by changing the model's prior belief about task performance.

The mechanism behind under- and overconfidence.

The last decade has seen an increase in studies investigating how the sense of confidence can be explained within models of decision making. In these endeavours, individual differences in the overall level of confidence are often treated as a nuisance and are typically not the primary focus of investigation. Instead, most prior work has focussed on measures that quantify the sensitivity of decision confidence (Maniscalco & Lau, 2012; Fleming & Lau, 2014). Much less attention has been devoted to the computational mechanisms underlying confidence biases. For example, in signal detection theory, an influential framework often used to quantify the sensitivity of decision confidence, biases in confidence can easily be modelled by changing the criteria that dissociate high from low confidence (Rahnev, 2021). However, this is merely descriptive and does not provide us with fundamental insight into *why* different people have different confidence criteria. To answer this question, in the current work we relied on dynamic probabilistic models that explain confidence as the probability of being correct given time and evidence (Kiani & Shadlen, 2009; Zylberberg et al., 2016). Although it has been suggested previously that priors in these models might be an important factor to understand deviations in the computation of confidence (Moreno-Bote, 2010; Drugowitsch et al., 2014; Khalvati et al., 2021), empirical evidence for this claim has been lacking so far. Here, we provide the first empirical demonstration that inducing under- and overconfidence by means of changes in prior beliefs can be readily accounted for within dynamic probabilistic models. This process of changing the mapping between evidence and probability correct is demonstrated in Figure 1, where each scenario generates a different heatmap illustrating this mapping. Although our prior belief induction might come across as rather artificial, it is not hard to imagine how this process might operate in real life. Spontaneous differential exposure to feedback (cf. Experiment 1) or mostly engaging in difficult versus easy tasks (cf. Experiment 2) will result in differences in prior beliefs about task ability and hence confidence ratings.

An important aspect of our modelling approach is that rather than having fitted our model to the empirically observed confidence judgments, we only exposed the model to the same feedback that participants were presented with. Thus, our model was effectively blind to the empirical confidence judgments. Nevertheless, it captured the influence of prior beliefs on confidence. Although one might argue that the effect seen in the empirical data is weaker compared to the model predictions, an important consideration here is that our model starts from a blank slate (i.e. without any prior knowledge or preconceptions) and then builds its prior beliefs entirely based on the feedback. Real participants, however, likely come to the experiment with pre-existing prior beliefs about their ability to perform well on the task, and our experimental manipulations ride on top of these preconceptions. As such, a fruitful approach for future work will be to unravel how our experimental manipulations interact with pre-existing beliefs about task performance. As a second note, our model assumed perfect integration of the feedback. This might be different from real humans, who likely display leaky integration of feedback. Nevertheless, despite these simplifications, our dynamic probabilistic model captured the empirical findings by changing the prior mapping between the amount of evidence and the probability of making a correct choice.

The interplay between "local" and "global" confidence.

The findings reported in the current work address the computation of so-called "local" decision confidence. In recent years, interest is growing into the computation of "global" decision confidence, that is confidence at the task level (often referred to as global performance estimates; Rouault et al. 2019; Lee et al., 2021). Global confidence is the general, subjective feeling that subjects have about their ability to perform a task, spanning a broader timescale than trial-specific local confidence. From this description, it becomes clear that global confidence bears close resemblance to the concept of prior beliefs discussed in the current work. Following this logic, our findings thus suggest a direct influence of prior beliefs about task performance (global confidence) on how we actually believe we are doing on individual trials (local confidence). Interestingly, it has been shown that in the absence of trial-by-trial feedback subjects compute global confidence by integrating local confidence judgements (Rouault et al., 2019). Thus, there also seems to be a direct influence of local confidence onto global confidence, revealing an intriguing interplay between global and local confidence. Therefore, it could be that causally inducing prior beliefs might have a long-lasting, self-sustaining effect on local confidence by means of a self-sustaining loop between local and global confidence. Our current data already suggest that the effect of the prior beliefs manipulation is not limited to the first few testing trials only: Even in the third and last testing block (72 trials), the effect was still visible, suggesting a long-lasting effect on confidence instead of a mere temporary boost or lapse in self-confidence specific to the training only. Future work might address whether the persistent nature of prior beliefs on local decision confidence is indeed mediated by global confidence. As impaired confidence has been linked to a variety of psychiatric symptoms (Hoven et al., 2019), uncovering the mechanisms behind these persistent biases could provide important new insights for the clinical field. For example, the self-sustaining nature of our prior beliefs manipulation could potentially provide a new tool to aid in restoring impaired confidence estimation in individuals with psychiatric disorders.

Dissociations between accuracy and confidence.

In recent years, an increasing number of studies have reported dissociations between confidence and accuracy (Vaghi et al. 2017; Rahnev, 2021). For example, it has been shown that whereas choices are equally informed by choice-relevant and choice-irrelevant information, decision confidence has been found to mostly reflect variation in choice-relevant information ("positive evidence bias"; Maniscalco et al., 2016; Peters et al., 2017; Koizumi et al., 2015; Zylberberg et al., 2012). In a similar vein, it has been reported that variance has a more profound effect on confidence than it has on decisions (Desender et al. 2018; Boldt et al. 2017; Spence et al. 2016; de Gardelle & Mamassian, 2015). Importantly, these observations are often treated as evidence for the existence of a metacognitive module existing separately from the decision making circuitry (Fleming & Daw, 2016; Mamassian & de Gardelle, 2021). In the current work, we reported a clear dissociation between accuracy and confidence, with only the latter being influenced by our manipulations of prior beliefs. Importantly, different from earlier work, our interpretation of these findings does not require a separate metacognitive module. Instead, we were able to explain decision confidence within the decision circuitry by simply changing the prior beliefs within this framework (see also Khalvati et al., 2021). Thus, our findings add to the ongoing debate about the need for a separate metacognitive module to explain dissociations between accuracy and decision confidence.

Counterfactual confidence.

One interesting discussion point is the extent to which participants were aware that they were reporting different levels of confidence for the same levels of evidence. Especially in Experiment 1, some participants indicated in the post-experiment debriefing that they were aware of giving lower confidence ratings after receiving negative feedback about their performance. Although these answers were given only after participants were informed about the true nature of the feedback, this nevertheless raises the intriguing question whether participants immediately computed the level of confidence that they eventually reported (i.e. modulated by prior beliefs), or whether instead they initially computed the "unbiased" probability of being correct and afterwards lowered/increased this rating depending on the feedback seen earlier. The latter option would imply that participants do have an unbiased representation of confidence at their disposal, which could be used for alternative purposes. Although such a dissociation might sound unlikely, previous work has shown evidence that participants compute so-called "counterfactual" confidence which they use to update their response bias (Zylberberg et al. 2018). Moreover, in social contexts it is known that people can sometimes feel (not) very confident but for social reasons decide to report a (higher) lower level of confidence (Bang et al. 2022).

Conclusion

In this work, we demonstrated that a manipulation of prior beliefs in task performance, either via comparative feedback or via changes in task difficulty, has a direct causal influence on subsequent ratings of decision confidence. These findings were well accounted for within a dynamic probabilistic model of decision confidence by changing the prior probabilities. Our findings provide a mechanistic understanding of the existence of under- and overconfidence.

Acknowledgments

This research was supported by a Franqui start-up grant (PXF-D8830) and a Research Council-Flanders grant (G0B0521N). All raw data and analysis code are openly available at [insert link upon publication].

Method

Participants

Fifty participants (eight men, one third gender, age: M = 19, SD = 4.9, range 17–52) took part in Experiment 1. Fifty participants (five men, age: M = 18.5, SD = 1, range 17–22) in Experiment 2. Due to chance level performance in at least one of the tasks, we removed two participants from Experiment 1 and three from Experiment 2. All participants participated in return for course credit and read and signed a written informed consent at the start of the experiment. All procedures were approved by the local ethics committee.

Stimuli and apparatus

Both experiments were conducted on a 22-inch DELL monitor with a 60 Hz refresh rate, using PsychoPy3 (Peirce et al., 2019). All stimuli were presented on a black background centred around the middle of the screen (radius 2.49° visual arc). Stimuli for the dot number task (white dots) were presented in two equally sized boxes (height 20°, width 18°) at an equal distance from the centre of the screen. Stimuli for the XO-task (white X's and O's) and dot colour task (red and blue dots) were presented in one box (height 22°, width 22°), centred around the fixation point.

Procedure

General. In both experiments, participants completed three decision-making tasks: a dot colour task, a dot number task and an XO-task (see **Figure 2**). Each task started with 120 training trials. In Experiment 1, participants were presented performance feedback every 24 trials, while in Experiment 2, feedback was given on every trial. After the training phase of a task, a test phase of 216 trials followed during which no feedback was provided, but instead participants indicated their level of confidence after each choice. For all tasks, a trial started with a fixation cross that was presented for 500 ms, after which the stimulus appeared for 200 ms or until a response was given. Participants indicated their choice using the C or N key using the thumbs of both hands. There was no time limit for responses. On test trials, participants additionally rated their confidence after each choice on a 6-point scale, labelled 'certainly wrong', 'probably wrong', 'maybe wrong', 'maybe correct', 'probably correct', and 'certainly correct' (reversed order for half the participants). Confidence was indicated using the 1, 2, 3, 8, 9 and 0 keys with the ring, middle and index fingers of both hands. There was no response limit for indicating confidence.

For each task, there were three levels of stimulus difficulty (easy, average or difficult). Stimulus properties for Experiment 1 were decided based on the results of a small pilot study (N=5). For Experiment 2, stimulus properties were revised based on the results of Experiment 1 in order to achieve a better matching of accuracy between tasks. Stimulus dependencies for each task can be found in **Table 1**.

		Dot colour task *Number of dominant colour dots. The total number of dots was always 80.*		Dot number task *Number of dots in the variable field. The reference field always contained 50 dots.*		XO-task *Number of dominant letters. The total number of letters was always 80.*	
		EXP 1	EXP 2	EXP 1	EXP 2	EXP 1	EXP 2
Difficulty level	Easy	61 – 65	61 – 65	+ or - 21 – 25	+ or - 21 – 25	70 – 75	70 – 74
	Average	51 – 55	46 – 50	+ or - 11 – 15	+ or - 11 – 15	51 – 55	53 – 57
	Difficult	41 – 45	41 – 45	+ or - 1 – 5	+ or - 1 – 5	41 – 45	42 – 46

Table 1. *Stimulus properties for each difficulty level, task and experiment.*

Dot colour task. On each trial, participants decided whether a field contained more (static) blue or red dots. The total number of dots was always 80, with differing proportions of red or blue dots depending on the difficulty condition. The position of dots was randomly generated on each trial.

Dot number task. On each trial, two fields were presented, one of which contained 50 dots and the other more or less than 50 dots. Participants decided which of the two fields contained the largest number of dots. The exact number of dots in the variable field differed depending on the difficulty condition. The position of dots was randomly generated on each trial.

XO-task. On each trial, participants decided whether a field contained more X's or O's. The total number of X's and O's was always 80, with differing proportions of X's or O's depending on the difficulty condition. The position of the letters was randomly generated on each trial.

Experiment 1: Prior belief induction in the comparative feedback experiment. In Experiment 1, prior beliefs about the ability to correctly perform the task were manipulated by means of fake performance feedback during the training phase. Participants were told that their feedback score was indicative of their performance (accuracy and reaction time) on the preceding trials relative to the performance of other participants that took part previously. Unknown to participants, feedback was predetermined to be either good, average or bad and feedback scores were randomly sampled according to the feedback condition. Each participant received good feedback on one task (inducing prior beliefs of low task performance), average feedback on another task, and bad feedback on a third task (inducing prior beliefs of high task performance; order and mapping with tasks counterbalanced between participants). For each task, participants received feedback after every 24 training trials, amounting to 5 feedback presentations per task. Feedback scores were pseudo-randomly generated on each feedback presentation and ranged between 5 and 30% in the negative feedback condition, between 37 and 62% in the average feedback condition and between 70 and 95% in the positive feedback condition. To increase credibility of the negative feedback, the second out of the five feedback screens showed average feedback (ranging between 32% and 36%, labelled as average). Likewise, the second out of five feedback screens in the positive feedback condition showed average feedback (ranging between 63% and 67%, labelled as average).

At the top of feedback screens, a verbal indication of the participant's score was presented: depending on the score, "Good performance:" in green, "Average performance:" in white or "Bad performance:" in red. After the colon, the score itself was presented in the same colour as the verbal indication. In the middle of the feedback screen, the participant's score was indicated in a visual way. A vertically oriented rectangle with no fill colour was presented, with the bottom line marked "worst performance", the top line marked "best performance" and a midline marked "average performance". The participant's score was used to colour the same percentage of the rectangle's total surface (starting at the bottom) in red (bad performance), white (average performance) or green (good performance) (see Figure 2).

Experiment 2: Prior belief induction via task difficulty. In Experiment 2, prior beliefs about the ability to correctly perform the task were induced by manipulating the difficulty of the task during the training phase in three levels. Contrary to Experiment 1, participants received genuine feedback on every trial: Each correct choice was followed by the word "Correct!" and each incorrect choice by "Wrong!". Each participant completed one task with a training phase consisting of only easy trials (inducing positive prior beliefs about task ability), another with a training phase of all average trials (inducing average prior beliefs), and another with a training phase of all difficult trials (inducing negative prior beliefs).

Statistical analyses

Data were analysed using mixed effects models. We started from models including the fixed factors and their interaction(s), as well as a random intercept for each participant. These models were then extended by adding random slopes, only when this significantly improved model fit. Confidence ratings and reaction times were analysed with linear mixed effects models, for which we report *F* statistics and the degrees of freedom as estimated by Satterthwaite's approximation. Accuracy was analysed using a generalised linear mixed model, for which we report X^2 statistics. All model fit analyses were done using the lmerTest package (Kuznetsova, Brockhoff, & Christensen,2017) in RStudio (RStudio Team, 2019).

Computational model

Bounded accumulation

We modelled the data using the drift diffusion model (DDM), which is a popular variant of the wider class of accumulation-to-bound models. In the DDM, noisy evidence is accumulated, the strength of which is controlled by a drift rate *v*, until one of two boundaries *a* or -*a* is reached. Non-decision components are captured by a non-decision time *ter* parameter. To simulate data from the model, random walks were used as a discrete approximation of the continuous diffusion process of the drift diffusion model (Ratcliff & Tuerlinckx, 2002). Each simulated random walk process started at z*a (here, z was an unbiased starting point of 0) and stopped once the accumulated evidence reached either -*a* or *a*. At each time step τ, accumulated evidence increased by Δ with probability p and went down by Δ with probability 1-p. Both p and Δ are given in Eq. (1):

$$\Delta = \sigma\sqrt{\tau}$$
$$p = \frac{1}{2}\left(1 + \frac{\mu\sqrt{\tau}}{\sigma}\right) \quad (1)$$

Drift rate is given by μ and within-trial variability is given by σ. In all simulations, τ was set to 1 ms, and σ was fixed to .1.

Accounting for prior beliefs

Within this model confidence is given by mapping the accumulated evidence, the reaction time and the choice on a 2-dimensional heatmap (as shown in Figure 1) representing the probability of being correct for any given evidence level, time and choice. To allow for the fact that confidence judgments are given later in time than choices, we allowed for additional post-decision evidence accumulation following boundary crossing before quantifying the probability correct (Desender et al., 2021). The duration of the post-decision evidence accumulation process was set to the median confidence reaction time observed during the testing phase for each subject. The heat maps were constructed by generating 200.000 random walks without absorbing bounds, using a given drift rate μ > 0 and its opposite -μ. The probability of being correct (i.e. average accuracy) for each combination of reaction time, evidence and choice was defined as the proportion of paths generated by μ (or -μ depending on the choice) among the total number of paths that crossed this reaction time by evidence coordinate. An important aspect is that these probability maps depend on the actual drift rate that is used to generate them; when generating heat maps with high versus low drift rates the probability of being correct will be high versus low (i.e. because high drift rates are associated with a higher accuracy). To model prior beliefs, we assumed that the drift rate parameter controlling the shape of the heat map can be different from the drift rate parameter controlling objective performance. To avoid confusion, we refer to the former as the prior belief parameter (Vs) and the latter as the drift rate (v).

We estimated Vs for each participant and each prior belief condition by fitting it to the feedback received by participants in the training phase. Since feedback was only given at the end of blocks of 24 trials, we first equally assigned the feedback value presented at the end of a block to every trial within that block. Then, we simulated predictions of the observed trials in the training phase from DDM parameters fitted to the training data in order to have access to the amount of accumulated evidence. The total number of simulated trials was equal to 10 times the number of observed trials. Confidence predictions for those simulated trials were next computed by reading the mapping obtained from heatmaps generated by Vs. The cost function was determined by the mean squared error between observed feedback and predicted confidence Eq (2):

$$\mathrm{M}SE = \frac{\sum_{i}^{Ntrial} \sum_{j}^{Nrep} (oFB_i - pCJ_{ij})^2}{Ntrial * Nrep} \quad (2)$$

With oFB_i corresponding to the feedback received at trial i and pCJ_{ij} corresponding to the confidence predicted for the jth simulation of trial i. Each observed trial's feedback was compared multiple times to predictions to account for the stochastic nature of the DDM. Since generating a heatmap is computationally costly, we pre-generated 500 heatmaps from values of Vs ranging from 0 to .5. In the case of heatmaps generated by large Vs, data points were missing at the low evidence/high reaction time part of the heatmaps. Median confidence over lower reaction times was used to fill the missing values in higher reaction times at each evidence level. The cost function was then computed for each pre-generated heatmap. The estimated Vs for each participant and prior belief condition was therefore equal to the one that generated the heatmap with the minimum MSE.

DDM fitting

For the training data of Experiment 1 as well as the testing data in both experiments, we fitted 5 DDM parameters to the accuracy and reaction time data: 3 drift rates (*v*; one for each trial difficulty level), the decision boundary (*a*) and the non-decision time (*Ter*). Since only one trial difficulty was presented per task in the training phase of Experiment 2, only one drift rate was fitted to the behaviour data. To estimate these parameters, we implemented quantile optimization. Specifically, we computed the proportion of trials in quantiles .1, .3, .5, .7 and .9 of reaction time, separately for corrects and errors. We then used a differential evolution algorithm, as implemented in the DEoptim R package (Mullen et al., 2011), to estimate the free parameters by minimising the sum of squared error function shown in Eq (3):

$$SSE = \sum_i (oRT_i - pRT_i)^2 \tag{3}$$

With oRT_i and pRT_i corresponding to the proportion of observed and predicted responses in quantile i, respectively, separately calculated for correct and error reaction times. Model fitting was done separately for each participant, phase (training vs testing), and experimental manipulation. The population size for the differential evolution algorithm was set to 10 times the number of free parameters, as recommended in Price et al., 2006. Predictions from the model were generated by simulating 5000 random walk paths for each drift rate to be fitted.

References

Ais, J., Zyllerberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition, 146*(1), 377–386. http://dx.doi.org/10.1016/j.cognition.2015.10.006

Bang, D., Moran, R., Daw, N. D., & Fleming, S. M. (2022). Neurocomputational mechanisms of confidence in self and others. *BioRxiv*, 2021.03.05.434065. https://doi.org/10.1101/2021.03.05.434065

Baranski, J.V., & Petrusic, W.M. (1994). The calibration and resolution of confidence in perceptual judgments. Perception & Psychophysics 55(4), 412–428. https://doi.org/10.3758/BF03205299

Belmi, P., & Neale, M., Reiff, D., & Ulfe, R. (2020). The Social Advantage of Miscalibrated Individuals: The Relationship Between Social Class and Overconfidence and Its Implications for Class-Based Inequality. *Journal of Personality and Social Psychology, 118*(2), 254 –282. https://doi.org/10.1037/pspi0000187

Boldt, A., de Gardelle, V., & Yeung, N. (2017). The impact of evidence reliability on sensitivity and bias in decision confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(8), 1520–1531. https://doi.org/10.1037/XHP0000404

de Gardelle, V., & Mamassian, P. (2015). Weighting mean and variability during confidence judgments. *PLoS ONE*, *10*(3). https://doi.org/10.1371/JOURNAL.PONE.0120870

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective Confidence Predicts Information Seeking in Decision Making. *Psychological Science*, *29*(5), 761–778. https://doi.org/10.1177/0956797617744771

Desender, K., Donner, T. H., & Verguts, T. (2021). Dynamic expressions of confidence within an evidence accumulation framework. *Cognition, 207*(104522). https://doi.org/10.1016/j.cognition.2020.104522

Drugowitsch, J., Moreno-Bote, R., & Pouget, A. (2014). Relation between Belief and Performance in Perceptual Decision Making. *PLOS ONE, 9*(5), e96511. https://doi.org/10.1371/JOURNAL.PONE.0096511

Fleming, S. M., & Daw, N. D. (2016). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91. https://doi.org/10.1037/REV0000045

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*(443). https://doi.org/10.3389/fnhum.2014.00443

Hoven, M., Lebreton, M., Engelmann, J. B., Denys, D., Luigjes, J., & van Holst, R. J. (2019). Abnormalities of confidence in psychiatry: an overview and future perspectives. *Translational Psychiatry, 9*(1), 1–18. https://doi.org/10.1038/s41398-019-0602-7

Khalvati, K., Kiani, R., & Rao, R. P. N. (2021). Bayesian inference with incomplete knowledge explains perceptual confidence and its deviations from accuracy. *Nature Communications*, *12*(1), 1–16. https://doi.org/10.1038/s41467-021-25419-4

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*(5928), 759–764. https://doi.org/10.1126/science.1169405

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron, 84*(6), 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015

Koizumi, A., Maniscalco, B., & Lau, H. (2015). Does perceptual confidence facilitate cognitive control? *Attention, Perception, and Psychophysics*, *77*(4), 1295–1306. https://doi.org/10.3758/s13414-015-0843-3

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lee, A. L. F., de Gardelle, V., & Mamassian, P. (2021). Global visual confidence. *Psychonomic Bulletin and Review*, *28*(4), 1233–1242. https://doi.org/10.3758/S13423-020-01869-7

Lyons, B. A., Montgomery, J. M., Guess, A. M., Nyhan, B., & Reifler, J. (2021). Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(23). https://doi.org/10.1073/pnas.2019527118

Mamassian, P., & de Gardelle, V. (2021). Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychological Review*. https://doi.org/10.1037/REV0000312

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

Maniscalco, B., Peters, M. A. K., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception & Psychophysics*, *78*(3), 923–937. https://doi.org/10.3758/S13414-016-1059-X

Moreno-Bote, R. (2010). Decision confidence and uncertainty in diffusion models with partially correlated neuronal integrators. *Neural computation, 22*(7), 1786–1811. https://doi.org/10.1162/neco.2010.12-08-930

Moses-Payne, M. E., Rollwage, M., Fleming, S. M., & Roiser, J. P. (2019). Postdecision Evidence Integration and Depressive Symptoms. *Frontiers in Psychiatry, 10*, 639. https://doi.org/10.3389/fpsyt.2019.00639

Mullen, K. M., Ardia, D., Gil, D. L., Windover, D., & Cline, J. (2011). DEoptim: An R package for global optimization by differential evolution. *Journal of Statistical Software, 40*(6), 1–26. https://doi.org/10.18637/jss.v040.i06

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*, 195–203. https://doi.org/10.3758/s13428-018-01193-y

Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., Doyle, W., Kuzniecky, R., Devinsky, O., Halgren, E., & Lau, H. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*, *1*(7), 1–8. https://doi.org/10.1038/s41562-017-0139

Petruzzello, S. J., & Corbin, C. B. (1988). The effects of performance feedback on female self-confidence. *Journal of Sport & Exercise Psychology, 10*(2), 174–183. https://doi.org/10.1123/jsep.10.2.174

Price, K. V., Storn, R.M., & Lampinen, J. A. (2006). Differential Evolution: A Practical Approach to Global Optimization. Springer-Verlag, Berlin.

Rahnev, D. (2021). A robust confidence–accuracy dissociation via criterion attraction. *Neuroscience of Consciousness*, *7*(1), 1–11. https://doi.org/10.1093/nc/niab039

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9*, 438–481. https://doi.org/10.3758/BF03196302

Ratcliff, R., McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation, 20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Rollwage, M., Dolan, R. J., & Fleming, S. M. (2018). Metacognitive Failure as a Feature of Those Holding Radical Beliefs. *Current Biology*, *28*(24), 4014-4021. https://doi.org/10.1016/J.CUB.2018.10.053

Rouault, M., Dayan, P., & Fleming, S.M. (2019). Forming global estimates of self-performance from local confidence. *Nat Commun, 10*(1), 1141. https://doi.org/10.1038/s41467-019-09075-3

RStudio Team (2019). *RStudio: Integrated Development for R.* Boston, MA: RStudio Inc. URL: http://www.rstudio.com/

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron*, *90*(3), 499–506. https://doi.org/10.1016/j.neuron.2016.03.025

Saposnik, G., Redelmeier, D., Ruff, C. C., & Tobler, P. N. (2016). Cognitive biases associated with medical decisions: a systematic review. *BMC Medical Informatics and Decision Making*, *16*(1), 1–14. https://doi.org/10.1186/s12911-016-0377-1

Schmack, K., Bosc, M., Ott, T., Sturgill, J. F., & Kepecs, A. (2021). Striatal dopamine mediates hallucination-like perception in mice. *Science, 372*(6537), eabf4740. https://doi.org/10.1126/science.abf4740

Schwardmann, P., & van der Weele, J. (2019). Deception and self-deception. *Nature human behaviour, 3*(10), 1055–1061. https://doi.org/10.1038/s41562-019-0666-7

Spence, M. L., Dux, P. E., & Arnold, D. H. (2016). Computations underlying confidence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(5), 671–682. https://doi.org/10.1037/XHP0000179

Vaghi, M. M., Luyckx, F., Sule, A., Fineberg, N. A., Robbins, T. W., & de Martino, B. (2017). Compulsivity Reveals a Novel Dissociation between Action and Confidence. *Neuron*, *96*(2), 348-354.e4. https://doi.org/10.1016/J.NEURON.2017.09.006

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*(SEP). https://doi.org/10.3389/FNINT.2012.00079

Zylberberg, A., Fetsch, C. R., & Shadlen, M. N. (2016). The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *ELife*, *5*. https://doi.org/10.7554/ELIFE.17688

Zylberberg, A., Wolpert, D. M., & Shadlen, M. N. (2018). Counterfactual Reasoning Underlies the Learning of Priors in Decision Making. *Neuron*, *99*(5), 1083-1097.e6. https://doi.org/10.1016/J.NEURON.2018.07.035