Full title:

Two sides of the same coin: monetary incentives concurrently improve and bias confidence judgments.

Running title:

How motivation impacts confidence accuracy.

Authors:

Maël Lebreton[1,2+] Shari Langdon[3,4], Matthijs J. Slieker[3,4], Jip S. Nooitgedacht[3,4], Anna E. Goudriaan[3,4], Damiaan Denys[4,5], Judy Luigjes[3,4*], Ruth J. van Holst[3,4*]

Affiliations:

[1]Amsterdam Brain and Cognition (ABC),

[2]CREED, Amsterdam School of Economics (ASE), Universiteit van Amsterdam, 1018 WB Amsterdam, the Netherlands.

[3]Amsterdam Institute for Addiction Research,

[4]Department of Psychiatry, Academic Medical Centre, 1100 DD Amsterdam, the Netherlands

[5]Netherlands Institute for Neuroscience, Institute of the Royal Netherlands Academy of Arts and Sciences, 1105 BA Amsterdam, The Netherlands

[+] To whom correspondence should be addressed (m.p.lebreton@uva.nl)

*shared last authorship.

Abstract:

Most decisions are accompanied by a feeling of confidence, i.e., a subjective estimate of the probability of being correct. Although confidence accuracy is critical, notably in high-stakes domains such as medical or financial decision-making, little is known about how incentive motivation influences this metacognitive judgment. In this article, we hypothesized that motivation can, paradoxically, deteriorate confidence accuracy. We designed an original incentive-compatible perceptual task to investigate the effects of monetary incentives on human confidence judgments. In line with classical theories of motivated cognition, our results first reveal that monetary incentives improve some aspects of confidence judgments. However, over three experiments and in line with our hypothesis, but unpredicted by normative or classical motivated cognition theories, we further show that incentives also robustly bias confidence reports: the perspective of potential gains (respectively losses) bias confidence upward (respectively downward), with potential detrimental consequences on confidence accuracy. Connecting our findings with recent models of confidence formation, we demonstrate that these two effects of incentives have dissociable signatures on how confidence builds on decision evidence. Altogether, these findings enrich current cognitive and evolutionary models of confidence, and may provide new hints about its healthy or pathological miscalibration.

Significance Statement:

Most decisions are accompanied by a feeling of confidence, i.e., a subjective estimate of the probability of being correct. Although achieving accurate confidence judgments is theoretically fundamental for individual decision-makers, miscalibrations such as overconfidence appear to be "widespread, stubborn, and costly" (D. Kahneman, in *Thinking Fast and Slow*. 2011). In this manuscript, we investigated the influence of incentive motivation on confidence accuracy in humans. In a series of behavioral experiments, we found that, although incentives can improve confidence accuracy, they also paradoxically systematically bias confidence judgments, thereby creating detrimental miscalibrations such as overconfidence. These findings have important implications for cognitive and evolutionary models of confidence, and may provide new hints about its healthy or pathological miscalibration.

Introduction:

Imagine you have to cross a road. It's dark and raining, the visibility is low. At some point, you estimate that there does not seem to be any danger, and decide to cross. However, because you feel quite unsure about crossing with such low visibility, you decide to check one last time, and spot a car coming at high speed to your direction. Luckily, you have enough time to withdraw from the street and avoid being hit by the car.

Just like in this example, most decisions in everyday life are accompanied by a subjective feeling of confidence emerging from the constant monitoring of our own thoughts and actions by metacognitive processes (1, 2). Formally, confidence is a decision maker's estimate of the probability –or belief- that her action, answer or statement is correct, based on the available evidence (3, 4). Although high confidence accuracy seems critical to monitor and re-evaluate previous decisions (5) or to arbitrate between different strategies (6, 7), converging evidence suggests that confidence judgments often significantly differ from the actual probability of being correct. Notably, we seem to often overestimate the probability of being correct, a phenomenon called overconfidence (8). This bias, potentially detrimental for the decision-maker or society, has been consistently reported in numerous domains and situations, from simple sensory psychophysics (9) or knowledge (10) tasks in the laboratory, to medical (11), financial, and managerial (12, 13) decision-making.

Altogether, the importance of confidence as a cognitive variable mitigating decision-making and the societal relevance of its miscalibration have considerably stimulated the search for the computational and neurobiological mechanisms subsuming confidence estimation (14, 15). Several studies have investigated factors which modulate or bias confidence estimation, and have notably established links between (over-)confidence and an elevated mood (16), absence of worry (17), emotional arousal (18, 19) or increased desires (20). Recently, functional neuroimaging studies also reported neural correlates of confidence in the ventromedial prefrontal cortex (21, 22), a brain region associated with the encoding of economic, motivational and affective values (23). Such an overlap in the neural correlates of confidence and values suggests that these variables also interact at the behavioral level. In practice, this hypothesis entails that a decision-maker reports higher confidence not only because she strongly believes to be correct or to perform better, but also because she is in a high expected- or experienced- value context. Although this values-confidence interaction could parsimoniously explain associations between positive affective-states and overconfidence (16–20), it could also be the cause of new critical motivational biases. We conjectured that incentivizing confidence accuracy paradoxically biases confidence reports: following expected values, we expect higher monetary incentives to bias confidence judgments upwards in a gain frame, and downward in a loss frame, despite the potentially detrimental consequences on the final payoff.

In order to test this hypothesis, we designed an original task, where participants had to first make a difficult perceptual decision, and then to judge the probability of their answer being correct, i.e., their confidence in their decision. Critically, we incentivized the truthful and accurate reporting of confidence using an adapted version of the Becker-DeGroot-Marschak (BDM) auction, a well-validated method from behavioral economics (24, 25). This incentivization was implemented after the perceptual choice, which made it possible to separately motivate the accuracy of confidence judgments without directly influencing the performance on the perceptual decision per se (**Fig. 1** and **Methods**). Briefly, the BDM auction mechanism considers participants' confidence reports as bets on the correctness of their answers, and implement trial-by-trial comparisons between these bets and random lotteries. Under utility maximization assumptions, this guarantees that participants maximize their earnings by reporting their most precise and truthful confidence estimation (26, 27). Experimental evidence and theories of motivated cognition from behavioral economics (28–30) and cognitive psychology (31) predict that higher stakes (irrespective of their valence, i.e., gain or loss) increase participants tendency to conform to rational model predictions, hence should improve confidence accuracy.

In three experiments, we systematically varied the monetary stakes magnitude and valence (gains or losses), used to incentivize confidence accuracy. Confirming our initial hypotheses and unpredicted by normative or motivated cognition theories, we show that monetary incentives bias confidence judgments: the prospect of potential gains increases confidence –in our case harming people's overall payoff-, while the prospect of potential losses decreases confidence. We further investigated the properties of this bias, and demonstrated that it is independent from the amount of evidence on which the choices are based.

Results:

Experiment 1. Twenty-four subjects participated in our first experiment. They performed four sessions of our confidence elicitation task (**Fig. 1, Table 1** and **Methods**): in each trial, participants briefly saw a pair of Gabor patches first, then had to indicate which one had the highest contrast, and finally had to indicate how confident they were in their answer (from 50 to 100%). Critically, the confidence judgment was incentivized: after the binary choice and before the confidence judgment, a monetary stake was displayed, which could be neutral (no incentive) or indicate the possibility to gain or lose 1 euro. Participants were explicitly instructed that they could maximize their chance to gain (respectively not lose) the stake by reporting their subjective probability of being correct as truthful as possible in the confidence judgment step, because a BDM incentivization mechanism (24, 25) determined the outcome of the trial. In addition to extensive instructions explaining the BDM procedure, participants gained direct experience with this procedure through a series of 24 training trials that did not count towards final payment.

Prior to the task, participants performed a calibration session, which we used to estimate the parameters of a logistic choice function linking individual choices (probability of choosing the left Gabor) and the contrast intensity difference (C_L-C_R) (see **Methods**). These parameters were then used to generate the main task stimuli, such that the decision situations spanned a pre-defined range of individual subjective difficulties. Results show a very good agreement between *ex ante* model choice predictions from this psychophysical model and actual subject choice behavior (R = .974 ± .003, t_{23} = 364, P = 9.06e-45; **Fig. 2.A**). In addition, both decision correctness and confidence are highly correlated with the amount of evidence, a measure which normalizes the contrast difference by their sum to adjust for saturation effects (correctness: logistic β = 7.46 ± .479, t_{23} = 15.6, P = 1.03e-13; and confidence linear β = .187 ± .018, t_{23} = 10.2, P = 5.09e-10; **Fig. 2.A**). These results confirm that our task is well calibrated, and that subjects' behavior conforms to a priori predictions.

We next assess three important properties of our confidence measure (**Fig. 2.B**). First, confidence is highly correlated with the probability of being correct (R = .736 ± .055, t_{23} = 13.2, P = 2.66e-12). Second, the link between confidence and evidence is positive for correct and negative for incorrect responses (correct: linear β = .193 ± .016, t_{23} = 11.8, P = 2.85e-11; and incorrect: linear β = -.288 ± .063, t_{23} = -4.58, P = 1.31e-4). Finally, the link between evidence and performance differs between high and low confidence trials (respectively, logistic β = 11.3 ± 1.00, and logistic β = 4.70 ± .637, difference: t_{34} = 5.45, P = 1.54e-5). These properties suggest that the confidence measure elicited in our task actually corresponds to subjects' estimated posterior probability of being correct (15).

We then turn to the original feature of our design: the incentivization of confidence accuracy (**Fig. 2.C**). Critically, monetary incentives have a significant impact on confidence ($F_{2,23}$= 8.26, P = 8.63e-4;). In line with our confidence-value interaction hypothesis, this translates into a monotonic increase of confidence from the loss to the neutral and gain conditions (R = .604 ± .090, t_{23} = 6.69, P = 8.02e-7), although post-hoc t-tests revealed that the prospect of winning 1 euro does not significantly differ from the neutral condition (**SI Results**). As expected from our task design, incentives have no effect on performance ($F_{2,23}$= .715, P = .494), which rules out the possibility that the effect of incentives on confidence is driven by an effect on performance.

In order to explore how the incentive effect on confidence judgments impacts confidence *accuracy*, we focused on two distinctive metacognitive metrics: discrimination and calibration (see **Methods** for formal definitions). Discrimination (or resolution) measures how confidence distinguishes between correct and incorrect responses, and is computed as the difference between the average confidence for correct answers and the average confidence for incorrect answers; the higher the discrimination, the more accurate the confidence judgments. On the other hand, calibration measures how averaged confidence matches averaged objective performance, and is computed as the difference between the averaged confidence and the averaged performance. Therefore, a calibration of zero signals high confidence accuracy, whereas a positive (respectively negative) calibration signals overconfidence (respectively underconfidence).

Our results first show that incentives affect discrimination ($F_{2,23}$ = 4.62, P = .0148): prospects of gains and losses lead to better discrimination than no-incentives (gain vs neutral: t_{23} = 2.33, P = .0288; loss vs neutral: t_{23} = 2.41, P = .0244). Because those results are in line with motivated cognition theories, we refer to this first effect as the *motivational* effect of incentives on confidence accuracy.

Second, incentives also significantly influence calibration ($F_{2,23}$= 6.73, P = 2.72 e-3); however, mirroring the effects on confidence, increasing incentive value lead to higher calibration, (R = .452 ± .115, t_{23} = 3.95, P = 6.43.e-4). Since participants are overconfident on average, calibration is thereby improved by loss prospects, but paradoxically worsened by gain prospects. We refer to this second effect as the *biasing* effect of incentives on confidence.

Because mechanistic models of confidence formation propose that it builds on perceptual evidence (15), we next explore how incentives modulate the relationship between those two quantities, for correct and incorrect answers (**Fig. 2.D**). We estimate linear regressions linking confidence and evidence for each incentive level, and assess whether incentives influence this linear relationship. Our results show a clear dissociation between the motivational and biasing effects of incentives on confidence in this model. On the one hand, the motivational effect is found in the slopes of those regressions (correct answers: $F_{2,23}$= 4.40, P = .0179; incorrect answers: $F_{2,23}$= 3.38, P = .0428): in both cases, gains and losses increase the linear relationship between confidence and evidence, compared to no incentives (**SI Results**). On the other hand, the biasing effect of incentives is found in the intercept of those regressions (correct answers: $F_{2,23}$= 7.21, P = 1.88e-3; incorrect answers: $F_{2,23}$= 6.35, P = 3.66e-3), paralleling the effect of incentives on confidence judgment (correct answers R = .498 ± .119, t_{23} = 4.20, P = 3.45.e-4; incorrect answers R = .420 ± .130, t_{23} = 4.20, P = 3.66.e-3). Therefore, while the motivational effect of incentives actually influences the way confidence is built from evidence, the biasing effect appeared to be a purely additive effect of incentives on confidence, unrelated to the amount of evidence.

Additional analyses assessed the possibility that incentives impact other critical features of our design, such as the predictive accuracy of our psychophysics model, or the participants' reaction time, but none of these effects are significant. (**SI Results**).

Overall this first set of results reveal and dissociate two concurrent effects of monetary incentives on confidence accuracy: a motivational effect, which improves discrimination with monetary stakes (gains or losses), and a biasing effect, which monotonically alters confidence with incentive values (i.e. upward for gains, and downward for losses), degrading calibration (i.e. increasing overconfidence) with increasing incentive values.

Experiment 2. If the value of monetary incentives actually interacts with confidence judgments -rather than simply creating a framing effect- then the magnitude of the incentives should modulate the observed effects. We invited thirty-five subjects to participate in a second task where incentives for confidence accuracy varied in both valence (gains and losses) and magnitude (1€ vs 10¢) (see **Table 1**). As a first sanity check, we replicate the validation of our experimental design together with the fundamental features of confidence (**SI Results**). A two way-ANOVA reveals significant effects of both valence (i.e. gains or losses) and magnitude (high vs low stakes) on confidence (valence: $F_{1,34}$= 24.7, P = 1.88e-5; magnitude: $F_{1,34}$= 3.40, P = .0740; valence*magnitude: $F_{1,34}$= 6.86, P = .0131; **Fig. 3.A**). This replicates our biasing effect, where incentives monotonically bias confidence reports (R = .658 ± .054, t_{34} = 12.2, P = 5.09.e-14). However, post-hoc t-tests show that the effect of magnitude is not significant in the gain domain (**SI Results**). This effect on confidence judgments percolates confidence accuracy, as a similar effect is found on calibration (R = .431 ± .073, t_{34} = 5.91, P = 1.11.e-6), similarly driven by the incentive valence (**SI Results**). Control analyses show no effects of incentives on performance (**SI Results**). Contradictory to the first experiment, discrimination was not significantly influenced by incentives, suggesting the motivational effect evidenced in the first experiment was due to the mere presence of incentives rather than the magnitude of those incentives (**SI Results**). Finally, we confirm that the biasing effect of incentives is independent from the amount of evidence, impacting the intercepts and not the slope of the linear relationship between evidence and confidence for both correct and incorrect answers (**Fig. 3.B** and **SI Results**).

The results from this second experiment have important implications. First, they replicate the biasing effect of incentives on confidence, and furthermore demonstrate that these effects depend on incentive magnitude. Additionally, they isolate this biasing effect from the motivational effects, supporting our proposed dissociation.

Experiment 3. While the biasing effect of incentives on confidence and calibration revealed in our first two experiments appeared robust and replicable, it seemed to be driven by the loss frame. A first hypothesis is that those biasing effects are purely restricted to the loss frame. However, an alternative hypothesis is that subjects are simply less sensitive to gains, as suggested by prospect theory (32). To dissociate between those two possibilities, we invited twenty-four subjects to participate in a final study which included higher stakes (2€) in both gain and loss frames (**Table 1**). As in the first two experiments, we replicate the validation of our experimental design, together with the fundamental features of confidence (**SI Results**). Once again, our results reveal a significant effect of incentives on confidence, which depends on both the valence and the magnitude of the stakes (valence: $F_{1,23}$= 18.8, P = 2.45e-4; magnitude: $F_{1,23}$= .770, P = .469; valence*magnitude: $F_{1,23}$= 6.52, P = 1.34e-5; **Fig. 3.C**). This translates into a parametric increase of confidence with incentives (R = .604 ± .082, t_{23} = 7.34, P = 1.80.e-7), with a significant effect of incentive magnitude in both gain and loss domains (**SI Results**). This result confirms our initial hypothesis: following expected values, higher incentives bias confidence judgments upwards in a gain frame, and downward in a loss frame, independent from the amount of evidence on which the decisions are based (**Fig. 3.D** and **SI Results**). This bias also affects confidence accuracy, i.e., calibration (R = .302 ± .100, t_{23} = 3.02, P = 6.04.e-3), although this is mostly driven by the valence effect (**SI Results**). Similar to the second experiment, though, the motivational effects on discrimination were not found, suggesting that they are mostly driven by the incentive versus no-incentive contrast (**SI Results**).

This last set of results replicates, for the third time, the biasing effects of incentives on confidence (see **SI Results** for a comparison of effect sizes between the three experiments), and confirms that monetary gains and losses both contribute to biasing confidence in perceptual decisions.

Discussion:

In this article, we investigated how monetary incentives influence confidence accuracy. To do so, we designed an original experimental setting which combined a perceptual decision task and a BDM auction procedure inspired from behavioral economics (24, 25). In addition to replicating important features of a recent model of confidence formation (15), we revealed and dissociated two effects of monetary incentives on confidence accuracy.

The first effect is a motivational effect: incentivizing confidence judgments improves discrimination. This means that high (respectively low) confidence is more closely associated with correct (respectively incorrect) decisions when confidence reports are incentivized, regardless of the valence of the incentive (gain or loss). This nicely extends a recent study reporting a similar effect of incentivization on discrimination, although limited to the gain domain (33). In addition, we show that this motivational effect is underpinned by a better integration of perceptual evidence in the confidence judgment when stakes increase, in line with theories of motivated cognition (28–31).

Although validating our initial hypothesis, the second effect of incentives on confidence accuracy is counter-intuitive: confidence judgments are parametrically biased by the value of the incentive. The prospect of gains increases confidence, while the prospect of losses decreases confidence. Because, people generally exhibited overconfidence in our experiment, the gain prospects detrimentally increased overconfidence (i.e. deteriorated calibration) while prospects of losses improved calibration. As opposed to the motivational effect, the biasing effect of incentive was purely additive, i.e., independent of the amount of evidence on which decisions and confidence judgments are based. While other studies linked (over)confidence with affective states such as elevated mood (16), absence of worry (17), emotional arousal (18, 19) or increased desires (20), the present study is, to our knowledge, the first to demonstrate this biasing effect of incentive values on confidence accuracy, to dissociate it from motivational effects, and to link those effects with models of confidence formation. One plausible interpretation for this effect is an affect-as-information effect: people use their momentary affective states as information in decision-making (34) which, in our case, means that they integrate the trial expected value into their confidence judgment. This interpretative framework could also connect the present findings to other results reported in the literature: in line with our hypotheses, Folke and colleagues (5) reported that confidence in value-based decisions is driven by the unsigned difference in value between options (i.e. decision evidence), but also biased by the summed value of the options.

In order to incentivize confidence reports, we used a mechanism inspired from BDM auction procedures (24, 25), sometimes referred to as reservation or matching probability. Contrary to other incentivization methods, such as the quadratic scoring rule (QSR), the BDM mechanism is still valid when subjects are not risk neural, and conveniently allowed us to manipulate the incentives on a trial-by-trial- basis (27, 33). Several studies have investigated the relative impact of different incentivization mechanisms on subjective probability judgments (confidence or belief), but with mixed results (see (27) for a review). Additional research should investigate how the value-confidence interaction impacts elicitation mechanisms which associate confidence levels with different payoffs, such as the QSR.

In this study, we only used relatively small monetary amounts as incentives; how the motivational and biasing effect of incentive scales when monetary stakes increase significantly remains an open question. Critically, higher stakes may also impact physiological arousal, which influence confidence and interoceptive abilities (19, 35). In general, the effects of incentives on confidence accuracy could also be mediated by inter-individual differences in metacognitive or interoceptive abilities (35, 36), and to incentive motivation sensitivity (37). Because our subject sample was mostly composed of university students, the generalization of those findings in the general population will have to be assessed in further studies.

Our results confirm that confidence judgments do not just represent rational estimates of the probability of being correct (4), but also integrate information and potential biases processed after a decision is made (38). The mere notion of confidence biases, notably overconfidence, and the actual conditions under which they can be observed sparked an intense debate in psychophysics (9, 39, 40) and evolutionary theories (41, 42). Critically, here, confidence accuracy was properly incentivized, hence deviations from perfect calibration can be appropriately interpreted as cognitive biases (42). The striking effects of incentive valence on confidence seem to make sense when considering evolutionary perspective: in natural settings, whereas overconfidence might pay off when prospects are potential gains, e.g., when claiming resources (41), a better calibration might be more appropriate when facing prospects of losses, e.g., death or severe injuries, given their potential dramatic consequences on reproductive chances. Interestingly, the observed valence difference in the effect of incentives magnitude –higher in the loss than in the gain domain- seem to mimic valence asymmetries observed in economic decision-making theories such as prospect theory (32).

How confidence is formed in the human brain and how neurophysiological constraints explain biases in confidence judgments remain critical research questions (4, 43). Although functional and structural neuroimaging studies initially linked confidence and metacognitive abilities to dorsal prefrontal regions (1), confidence activations were also recently reported in the ventro-medial prefrontal cortex (21, 22, 44), a region which has been consistently involved in motivation and value-based decision making (23). It is therefore possible that this region plays a role in the motivational and biasing effects of incentives on confidence, and constitutes the neurophysiological basis for the affect-as-information theory (34). However, this remains highly speculative and should be investigated in future neuroimaging studies.

Overall, our results suggest that investigating the interactions between incentive motivation and confidence judgments might provide valuable insights on the cause of confidence miscalibration in healthy and pathological settings.

Experimental Procedures

Subjects. All studies were approved by the local Ethics Committee of the University of Amsterdam Psychology Department. All subjects gave informed consent prior to partaking in the study. The subjects were recruited from the laboratory's participant database (https://www.lab.uva.nl/spt/). A total of 83 subjects took part in this study (see Table 1). They were compensated with a combination of a base amount (10€), and additional gains and/or losses from randomly selected trials (one per incentive condition per session for experiment 1, and one per incentive condition from one randomly selected session for experiments 2 and 3).

Tasks. All tasks were implemented using MATLAB® (MathWorks) and the COGENT toolbox (http://www.vislab.ucl.ac.uk/cogent.php). In all three experiments, trials shared the same basic steps (**Fig. 1.A**): after a brief fixation cross (750 ms) participants viewed a pair of Gabor patches displayed on both sides of a computer screen (150 ms), and judged which had the highest contrast (self-paced), by using the left or right arrow. They were thereafter presented with a monetary stake (1000 ms), and asked to report their confidence C in their answer on a scale from 50 to 100%, by moving a cursor with the left and right arrows and selecting their desired answer by pressing the spacebar (self-paced). The steps following the confidence rating, and the relation between the monetary stake, the confidence and the correctness of the answer were manipulated in two main versions of this task. In the ***Extended Version,*** at the trial level, the *lottery draw* step was separated in two smaller steps: first, a lottery number L was drawn in a uniform distribution between 50 and 100% and displayed as a scale under the confidence scale. After 1200 ms the scale with the highest number was highlighted for 1200 ms. Then, during the *resolution step*, if C happened to be higher than L, a clock was displayed for 750 ms together with the message "*Please wait*". Then, a *feedback* was displayed which depended on the correctness of the initial choice. Back at the *resolution step*, if the L happened to be higher than C, the lottery was implemented. A wheel of fortune, with a L% chance of losing was displayed, and played: the lottery arm spin for ~750 ms, and would end up in the winning (green) area with L% probability or in the losing (red) area with 1-L% probability. Then, a *feedback* informed whether the lottery was winning or losing.

Subject would win (gain frame) or not lose (loss frame) the incentive in case of a "winning" trial, and they would not win (gain frame) or lose (loss frame) the incentive in case of a "losing" trial. Thanks to this BDM auction procedure, the strategy to maximize one's earnings is to always report on the confidence scale one's subjective probability of being correct as truthfully and accurately as possible (**SI Text**). Subjects were explicitly instructed so. In the ***Short version***, the incentivization scheme was the same as in the ***Extended Version***, but part of it was run in the background. Basically, the lottery scale appeared, and the scale with the highest number was highlighted concomitantly (1200ms). Besides, the *resolution step* was omitted. Still, the complete feedback relative to the lottery and or the correctness of the answer was given to subjects in the *feedback* step. Analysis of Experiment 2 data showed that participants' behavior is not impacted by the different versions of the task (**SI Results**).

Stimuli & design: Participants initially performed a 144 trials calibration session (~5min), where they only performed the Gabor contrast discrimination task, without incentive or confidence measure (**Fig. 1.A**). During this calibration, the distribution of contrast difference (i.e. difficulty) was adapted every 12 trials following a staircase procedure, such that performance reached approximatively 70% correct.

The calibration data was used to estimate individual psychometric function:

$$p(\mathbf{ch_L}) = \ 1 + \exp(-\mu - \ \sigma \times (\mathbf{C_L} - \mathbf{C_R}))^{-1}$$

where $p(\mathbf{ch_L})$ is the probability of subjects choosing the left Gabor, and $\mathbf{C_L}$ and $\mathbf{C_R}$ are the contrast intensities of the left and right Gabors. In this formalization, μ quantifies subjects' bias toward choosing the left Gabor in the absence of evidence, and σ quantifies subjects' sensitivity to contrast difference. The estimated parameters (μ and σ) were used to generate stimuli for the confidence task, spanning defined difficulty levels (i.e. known $p(\mathbf{ch_L})$) for all incentives levels. After the first session of the confidence task, μ and σ were re-estimated for each session from the data of the preceding session (experiments 1 and 3), or from a new calibration session (experiment 2).

Statistics. All statistical analyses were performed with Matlab. All statistical analyses are based on second-level two-sided tests across subjects, using repeated measures n-way(s) ANOVAs with subjects as a random-effect, or (one-sample or paired) t-tests. Reported correlations were computed at the individual level, on binned or averaged data. Reported parameters from logistic and/or linear multiple regressions were computed at the individual level, on all-trial data, aggregated across sessions. In both cases, statistical tests are performed on the parameters at the population level, using t-tests or ANOVAs.

Metacognitive metrics. Calibration was computed as

$$\mathbf{C} = \frac{1}{n}\sum_{k=1}^{n} C_k - \frac{1}{n}\sum_{k=1}^{n} P_k$$

Where n is the total number of trials, Ck is the reported confidence at trial k, and Pk is the performance at trial k (1 for a correct answer and 0 for an incorrect anser);

Discrimination was computed as

$$\mathbf{C} = \frac{1}{n_c}\sum_{k_c=1}^{n_c} C_{k_c} - \frac{1}{n_i}\sum_{k_i=1}^{n_i} C_{k_i},$$

Where n_c (respectively n_i) is the number of correct (respectively incorrect answer).

Acknowledgments. ML, JL and RJvH were supported by Amsterdam Brain and Cognition (ABC) Talent Grants (Universiteit van Amsterdam). ML was additionally supported by an NWO Veni Fellowship, and the Bettencourt Schueller Fondation. AEG was supported by an NWO VIDI grant. The authors thank Stefano Palminteri, Frans van Winden, Jan B. Engelmann and Joel van der Weele for helpful discussions and comments on the manuscript, and Thomas A. Davison for checking the English.

References

1. Fleming SM, Dolan RJ (2012) The neural basis of metacognitive ability. *Phil Trans R Soc B* 367(1594):1338–1349.

2. Yeung N, Summerfield C (2012) Metacognition in human decision-making: confidence and error monitoring. *Philos Trans R Soc B Biol Sci* 367(1594):1310–1321.

3. Adams JK (1957) A Confidence Scale Defined in Terms of Expected Percentages. *Am J Psychol* 70(3):432–436.

4. Pouget A, Drugowitsch J, Kepecs A (2016) Confidence and certainty: distinct probabilistic quantities for different goals. *Nat Neurosci* 19(3):366–374.

5. Folke T, Jacobsen C, Fleming SM, Martino BD (2016) Explicit representation of confidence informs future value-based decisions. *Nat Hum Behav* 1:0002.

6. Donoso M, Collins AGE, Koechlin E (2014) Foundations of human reasoning in the prefrontal cortex. *Science* 344(6191):1481–1486.

7. Vinckier F, et al. (2016) Confidence and psychosis: a neuro-computational account of contingency learning disruption by NMDA blockade. *Mol Psychiatry* 21(7):946–955.

8. Lichtenstein S, Fischhoff B, Phillips LD (1982) Calibration of probabilities: the state of the art to 1980. *Judgment Under Uncertainty: Heuristics and Biases*, eds Kahneman D, Slovic P, Tversky A (Cambridge University Press, Cambridge, UK), pp 306–334.

9. Baranski JV, Petrusic WM (1994) The calibration and resolution of confidence in perceptual judgments. *Percept Psychophys* 55(4):412–428.

10. West RF, Stanovich KE (1997) The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. *Psychon Bull Rev* 4(3):387–392.

11. Berner ES, Graber ML (2008) Overconfidence as a Cause of Diagnostic Error in Medicine. *Am J Med* 121(5, Supplement):S2–S23.

12. Camerer C, Lovallo D (1999) Overconfidence and Excess Entry: An Experimental Approach. *Am Econ Rev* 89(1):306–318.

13. Malmendier U, Tate G (2005) CEO Overconfidence and Corporate Investment. *J Finance* 60(6):2661–2700.

14. Meyniel F, Schlunegger D, Dehaene S (2015) The Sense of Confidence during Probabilistic Learning: A Normative Account. *PLOS Comput Biol* 11(6):e1004305.

15. Sanders JI, Hangya B, Kepecs A (2016) Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron* 90(3):499–506.

16. Koellinger P, Treffers T (2015) Joy Leads to Overconfidence, and a Simple Countermeasure. *PLOS ONE* 10(12):e0143263.

17. Massoni S (2014) Emotion as a boost to metacognition: How worry enhances the quality of confidence. *Conscious Cogn* 29:189–198.

18. Jönsson FU, Olsson H, Olsson MJ (2005) Odor emotionality affects the confidence in odor naming. *Chem Senses* 30(1):29–35.

19. Allen M, et al. (2016) Unexpected arousal modulates the influence of sensory noise on confidence. *eLife* 5:e18103.

20. Giardini F, Coricelli G, Joffily M, Sirigu A (2008) Overconfidence in Predictions as an Effect of Desirability Bias. *Advances in Decision Making Under Risk and Uncertainty*, Theory and Decision Library., eds Abdellaoui PM, Hey PDJD (Springer Berlin Heidelberg), pp 163–180.

21. De Martino B, Fleming SM, Garrett N, Dolan RJ (2013) Confidence in value-based choice. *Nat Neurosci* 16(1):105–110.

22. Lebreton M, Abitbol R, Daunizeau J, Pessiglione M (2015) Automatic integration of confidence in the brain valuation signal. *Nat Neurosci* 18(8):1159–1167.

23. Levy DJ, Glimcher PW (2012) The root of all value: a neural common currency for choice. *Curr Opin Neurobiol* 22(6):1027–1038.

24. Becker GM, DeGroot MH, Marschak J (1964) Measuring Utility by a Single-Response Sequential Method. *Behav Sci* 9(3):226–232.

25. Ducharme WM, Donnell ML (1973) Intrasubject comparison of four response modes for "subjective probability" assessment. *Organ Behav Hum Perform* 10(1):108–117.

26. Schotter A, Trevino I (2014) Belief Elicitation in the Laboratory. *Annu Rev Econ* 6(1):103–128.

27. Schlag KH, Tremewan J, Weele JJ van der (2015) A penny for your thoughts: a survey of methods for eliciting beliefs. *Exp Econ* 18(3):457–490.

28. Smith VL, Walker JM (1993) Monetary Rewards and Decision Cost in Experimental Economics. *Econ Inq* 31(2):245–261.

29. Bonner SE, Sprinkle GB (2002) The effects of monetary incentives on effort and task performance: theories, evidence, and a framework for research. *Account Organ Soc* 27(4–5):303–345.

30. Wilcox NT (1993) Lottery Choice: Incentives, Complexity and Decision Time. *Econ J* 103(421):1397–1417.

31. Botvinick M, Braver T (2015) Motivation and Cognitive Control: From Behavior to Neural Mechanism. *Annu Rev Psychol* 66(1):83–113.

32. Tversky A, Kahneman D (1991) Loss Aversion in Riskless Choice: A Reference-Dependent Model. *Q J Econ* 106(4):1039–1061.

33. Hollard G, Massoni S, Vergnaud J-C (2015) In search of good probability assessors: an experimental comparison of elicitation rules for confidence judgments. *Theory Decis* 80(3):363–387.

34. Schwarz N, Clore GL (1983) Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *J Pers* 45(3):513–523.

35. Kandasamy N, et al. (2016) Interoceptive Ability Predicts Survival on a London Trading Floor. *Sci Rep* 6:32986.

36. Song C, et al. (2011) Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Conscious Cogn* 20(4):1787–1792.

37. Harsay HA, Cohen MX, Reneman L, Ridderinkhof KR (2011) How the aging brain translates motivational incentive into action: The role of individual differences in striato-cortical white matter pathways. *Dev Cogn Neurosci* 1(4):530–539.

38. Navajas J, Bahrami B, Latham PE (2016) Post-decisional accounts of biases in confidence. *Curr Opin Behav Sci* 11:55–60.

39. Olsson H, Winman A (1996) Underconfidence in sensory discrimination: The interaction between experimental setting and response strategies. *Percept Psychophys* 58(3):374–382.

40. Juslin P, Winman A, Olsson H (2000) Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychol Rev* 107(2):384–396.

41. Johnson DDP, Fowler JH (2011) The evolution of overconfidence. *Nature* 477(7364):317–320.

42. Marshall JAR, Trimmer PC, Houston AI, McNamara JM (2013) On evolutionary explanations of cognitive biases. *Trends Ecol Evol* 28(8):469–473.

43. Meyniel F, Sigman M, Mainen ZF (2015) Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron* 88(1):78–92.

44. Hebart MN, Schriever Y, Donner TH, Haynes J-D (2014) The Relationship between Perceptual Decision Variables and Confidence in the Human Brain. *Cereb Cortex*:bhu181.

	Exp. 1	Exp. 2	Exp. 3
M/F (total)	6/18 (24)	9/26 (35)	7/17 (24)
Age	23.8±6.00	24.8±5.43	24.3±7.79
Stakes (€)	0; ±1;	±0.1; ±1;	±0.1; ±1;±2
P(c=L)	50 ± [10:10:40] (×4)	50±[0:05:15] (×2)	50±[0:10:40] (×2)
Tasks	Short (×4)	Ext. (×1) + Short(×1)	Short (×3)

Table.1: Demographics and experimental design. P(c=L) indicates the level of difficulty (i.e probability of choosing the left Gabor) used to generate the stimuli. The number of times all levels of difficulty were repeated per incentive and per session is indicated between brackets. Tasks indicate which task version (short or extended) was offered to participants, and the number of session per task is indicated between brackets.

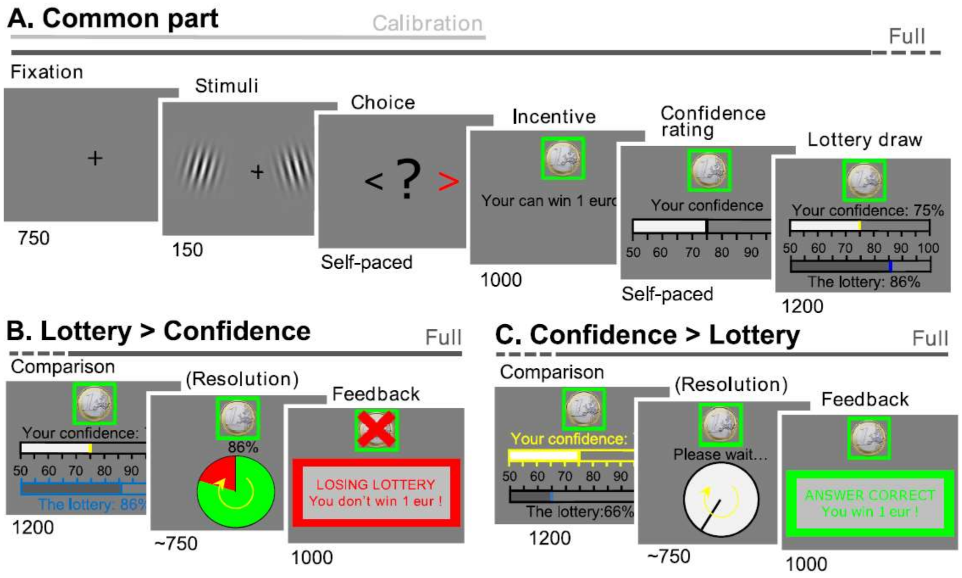

Fig. 1. Behavioral task.

Successive screens displayed in one trial are shown from left to right with durations in ms.

A. Common part. Participants viewed a couple of Gabor patches displayed on both sides of a computer screen, and judged which had the highest contrast. They were thereafter presented with a monetary stake (in a green frame for gain, grey for neutral and red for losses), and asked to report their confidence C in their answer on a scale from 50 to 100%. Then, a lottery number L was drawn in a uniform distribution between 50 and 100%, displayed as a scale under the confidence scale and the scale with the highest number was highlighted.

B. Lottery > Confidence. If the L > C, the lottery was implemented. A wheel of fortune, with a L% chance of losing was displayed, and played. Then, a *feedback* informed whether the lottery was winning or losing.

C. Confidence > Lottery. If C > L, a clock was displayed together with the message "*Please wait*", followed by a *feedback* which depended on the correctness of the initial choice.

Subject would win (gain frame) or not lose (loss frame) the incentive in case of a "winning" trial, and they would not win (gain frame) or lose (loss frame) the incentive in case of a "losing" trial.

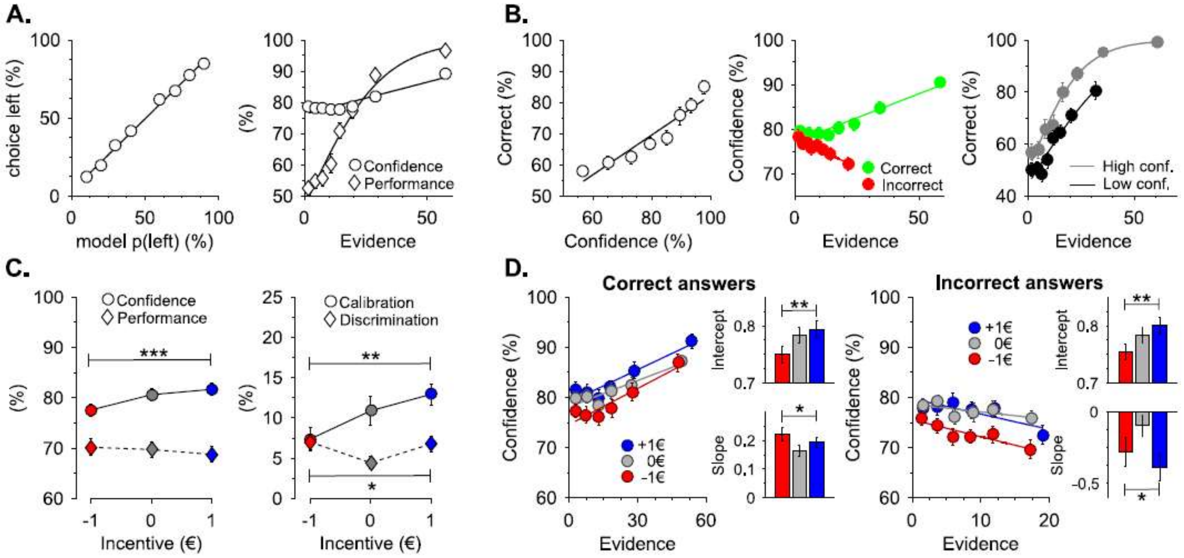

Fig. 2. Experiment 1.

A. General behavior. Left: Frequency of left Gabor choices, plotted against the *ex-ante* predictions from the psychometric model. Right: reported confidence (dots) and performance (diamonds) –i.e. % correct- as a function of evidence. Evidence is calculated as abs(C_L-C_R)./(C_L+C_R), where C_L and C_R are the contrast values of the left and right Gabor, respectively.

B. Statistical model of confidence. Left: observed performance (% correct choices) as a function of reported confidence. Middle: reported confidence as a function of evidence for correct (green) and incorrect (red) choices. Right: observed performance (% correct) as a function of evidence, for high (gray) and low (black) confidence trials.

C. Overall incentive effects. Left: reported confidence (dots) and performance (diamonds) –i.e. % correct- as a function of incentives. Right: computed calibration (dots) and discrimination (diamonds) as a function of incentives.

D. Linking incentives, evidence and confidence for correct (left) and incorrect (right) answers. In those two panels, the scatter plots display reported confidence as a function of evidence, for the different incentive levels. The histograms represent the intercepts (top) and slope (bottom) of this relationship, estimated at the individual level and averaged at the population level. In B and D, the solid line represents the best (linear or logistic) regression fit at the population level. Error bars indicate inter-subject standard errors of the mean. *: P<.05; ** P<.01; ***P<.001;

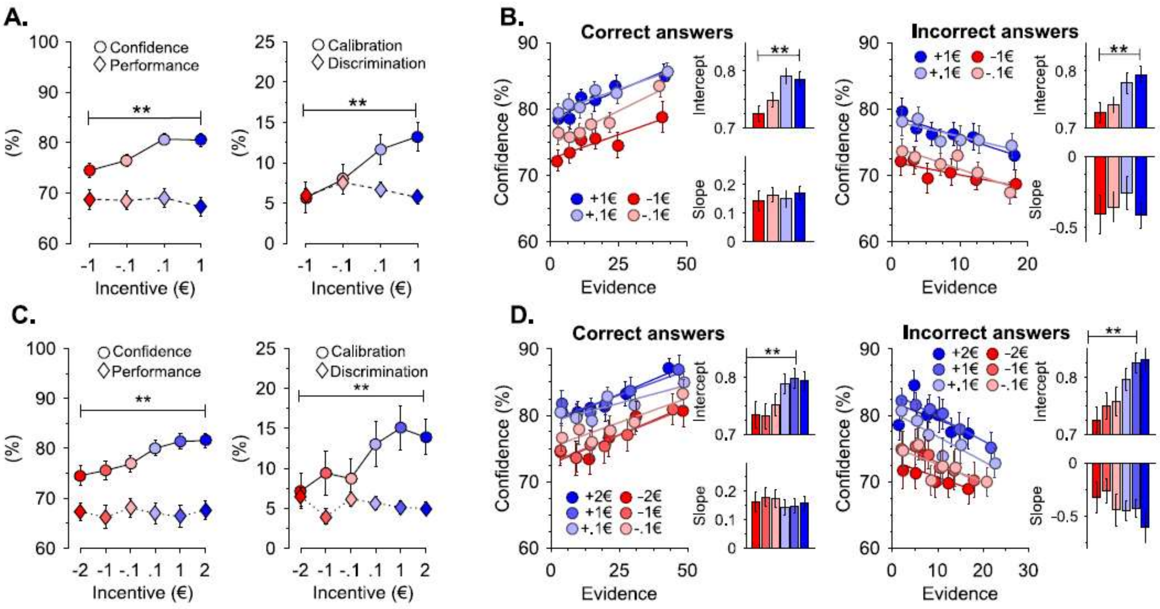

Fig. 3: Experiments 2 and 3.

These panels depict the results of experiment 2 (**A** and **B**) and 3 (**C** and **D**).

A and C. Overall incentive effects**.** Left: reported confidence (dots) and performance (diamonds) –i.e. % correct- as a function of incentives. Right: computed calibration (dots) and discrimination (diamonds) as a function of incentives.

B and D. Linking incentives, evidence and confidence for correct (left) and incorrect (right) answers. In those panels, the scatter plots display reported confidence as a function of evidence, for the different incentive levels. The histograms represent the intercepts (top) and slope (bottom) of this relationship, estimated at the individual level and averaged at the population level. In B and D, the solid line represents the best (linear) regression fit at the population level. Error bars indicate inter-subject standard errors of the mean. *: P<.05; ** P<.01; ***P<.001;