PSYCHOLOGY

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

Maël Lebreton,[1,2]* Shari Langdon,[3,4,5] Matthijs J. Slieker,[3,4] Jip S. Nooitgedacht,[3,4] Anna E. Goudriaan,[3,4,6] Damiaan Denys,[4,7] Ruth J. van Holst,[3,4†] Judy Luigjes[3,4†]

Decisions are accompanied by a feeling of confidence, that is, a belief about the decision being correct. Confidence accuracy is critical, notably in high-stakes situations such as medical or financial decision-making. We investigated how incentive motivation influences confidence accuracy by combining a perceptual task with a confidence incentivization mechanism. By varying the magnitude and valence (gains or losses) of monetary incentives, we orthogonalized their motivational and affective components. Corroborating theories of rational decision-making and motivation, our results first reveal that the motivational value of incentives improves aspects of confidence accuracy. However, in line with a value-confidence interaction hypothesis, we further show that the affective value of incentives concurrently biases confidence reports, thus degrading confidence accuracy. Finally, we demonstrate that the motivational and affective effects of incentives differentially affect how confidence builds on perceptual evidence. Together, these findings may provide new hints about confidence miscalibration in healthy or pathological contexts.

INTRODUCTION

In many situations, the ability to accurately assess the quality of our answers, actions, or statements is critical. Imagine analysts (for example, in the medical or financial domain) handing in independent recommendations on a case: It is crucial for the entity responsible for the final decision to survey as precisely as possible how confident each analyst is in his or her judgment to weigh their recommendations and come to the best final decision (*1*).

Confidence is formalized as the probability—or belief—that an action, answer, or statement is correct, based on the available evidence (*2*, *3*). Actually, 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 (*4*, *5*). Measuring confidence accuracy—that is, the quality of metacognitive judgments—is challenging (*6*–*8*), but confidence accuracy can consensually be split into a bias (or calibration) component measuring how confidence judgments differ from the overall probability of being correct and a sensitivity (or discrimination) component measuring how reliably confidence judgments can dissociate correct from incorrect answers (*6*, *7*).

Although high confidence accuracy seems critical to monitor and reevaluate previous decisions (*9*), to track changes in the environment (*10*), or to arbitrate between different strategies (*11*, *12*), converging evidence suggests that confidence judgments are significantly biased. Notably, we often overestimate the probability of being correct, a phenomenon called overconfidence (*13*). This bias, potentially detrimental for the decision-maker or society, has been consistently reported in numerous domains and situations from simple sensory psychophysics (*14*) or knowledge (*15*) tasks in the laboratory to medical (*16*), financial, and managerial (*17*, *18*) decision-making.

Back to our analyst example, standard theories of rational decision-making and motivation from behavioral economics (*19*–*21*) and cognitive psychology (*22*) advocate that properly incentivizing confidence accuracy (for example, with a financial bonus conditional on the precision of the estimation) should elicit less biased and more sensitive judgments. However, although this idea appears highly intuitive and is commonly applied, two lines of research suggest that it can actually have detrimental consequences on the quality of confidence judgments. The first line of research, encapsulated under the term "motivated cognition" or "motivated reasoning," has suggested that beliefs are influenced by individuals' desires (*23*–*25*). In other terms, individuals tend to estimate desirable events (like earning a bonus) to be more likely than undesirable ones, potentially leading to overconfidence (*26*). Studies have also established links between incidental psychological states such as elevated mood (*27*), absence of worry (*28*), or emotional arousal (*29*, *30*) and (over)confidence. The second line of research, leveraging functional neuroimaging, has recently reported neural correlates of confidence in the ventromedial prefrontal cortex (*31*, *32*), as well as in mesolimbic and striatal regions (*33*, *34*), a brain network associated with the encoding of economic, motivational, and affective values (*35*). 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 believes she is correct but also because she is in a high expected- or experienced-value context. This value-confidence interaction could explain associations between positive affective states and overconfidence (*26*–*30*), thereby underpinning biases in confidence judgments.

Here, we methodically investigated the interactions between incentive motivation and confidence in an attempt to explain features of human confidence accuracy. To do so, we designed a task where participants had to first make a difficult perceptual decision and then judge the probability of their answer being correct, that is, their confidence in their decision (Fig. 1A). To identify the critical features of the interactions between incentive motivation and confidence accuracy, the accuracy of confidence was incentivized with monetary prospects whose magnitude and valence were systematically varied

[1]Amsterdam Brain and Cognition, Universiteit van Amsterdam, 1018 WB Amsterdam, Netherlands. [2]Center for Research in Experimental Economics and Political Decision Making, Amsterdam School of Economics, Universiteit van Amsterdam, 1018 WB Amsterdam, Netherlands. [3]Amsterdam Institute for Addiction Research, Academic Medical Centre, 1100 DD Amsterdam, Netherlands. [4]Department of Psychiatry, Academic Medical Centre, 1100 DD Amsterdam, Netherlands. [5]Department of Pediatrics, Emma Kinderziekenhuis, Academic Medical Centre, 1100 DD Amsterdam, Netherlands. [6]Arkin Mental Health Care, 1070 AV Amsterdam, Netherlands. [7]Netherlands Institute for Neuroscience, Institute of the Royal Netherlands Academy of Arts and Sciences, 1105 BA Amsterdam, Netherlands.
*Corresponding author. Email: m.p.lebreton@uva.nl
†These authors shared last authorship.

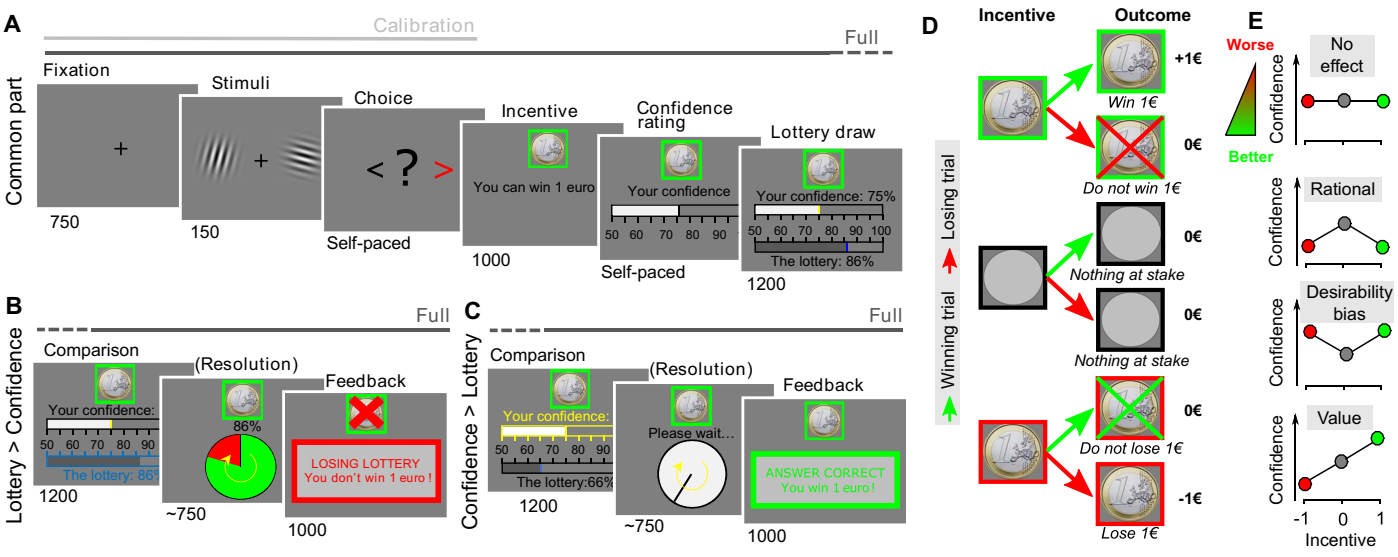

Fig. 1. Behavioral task and hypotheses. Successive screens displayed in one trial are shown from left to right with durations in milliseconds. (**A**) Behavioral task—Common part. Participants viewed a couple of Gabor patches displayed on both sides of a computer screen, and judged which had the highest contrast. They were then presented with a monetary stake (in a green frame for gain, gray 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**) Behavioral task—Lottery > Confidence. If *L* > *C*, then the lottery was implemented. A wheel of fortune, with an *L*% chance of winning, was displayed and played. Then, feedback informed whether the lottery resulted in a win or a loss. (**C**) Behavioral task—Confidence > Lottery. If *C* > *L*, then a clock was displayed together with the message "Please wait," followed by feedback that depended on the correctness of the initial choice. Subjects 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. (**D**) Behavioral task—Payoff matrix. Depending on the combination of a trial's offered incentive and the trial's final win or loss (regardless of whether the lottery or the correctness of the answer determined it), participants could receive various outcomes, from winning the proposed incentive to losing the proposed incentive. (**E**) Hypotheses. Expected biasing effects of incentives (−1€; 0€ or +1€) on confidence under different theoretical hypotheses. (Top) H0: No biasing effects of incentives. Participants are similarly overconfident in the three incentive conditions. (Middle top) H1: Rational decision-making. Under higher incentives, participants are more rational, that is, less biased. The absolute value of incentives therefore decreases confidence, if participants are generally overconfident. (Middle bottom) H2: Desirability bias. Participants are more inclined to believe that they are correct when higher incentives are at stake. The absolute value of incentives increases confidence. (Bottom) H3: Value-confidence interaction. The confidence judgment of participants is affected by the affective component of incentives. The net incentive value affects confidence.

(see Fig. 1 and Materials and Methods). This experimental manipulation elegantly orthogonalized the net incentive value (that is, the affective component of incentives, which can take both positive and negative values, indexed as V) and the absolute incentive value (that is, the motivational value of incentives, regardless of their valence, indexed as $|V|$). We used this experimental setup to investigate the effects of those two aspects of incentives on the two core components of confidence accuracy: bias and sensitivity.

Orthogonalizing the affective and motivational components of incentives enabled us to test three opposing predictions from three different theories anticipating effects of incentives on confidence bias (Fig. 1E). First, as outlined in the previous paragraph, standard theories of rational decision-making and motivation from behavioral economics (*19*–*21*) and cognitive psychology (*22*) predict that higher stakes increase participants' tendency to conform to rational model predictions and hence improve confidence accuracy regardless of the incentive valence. An increase in absolute incentive value should therefore increase confidence sensitivity and decrease confidence bias. In this case, we expect that if participants are generally biased toward overconfidence, then an increase in absolute incentive value should reduce this bias and therefore decrease confidence judgments. Second, motivated cognition theories (*23*)—that is, in the form of the desirability bias (*26*)—predict that participants should be more motivated to believe that they are correct when more money is at stake, irrespective of the valence (gain or loss). In this case, an increase in absolute incentive value should increase confidence judgments (and exaggerate the overconfidence bias). Finally, our value-confidence interaction hypothesis predicts that higher monetary incentives should bias confidence judgments upward in a gain frame, and downward in a loss frame, despite the potentially detrimental consequences on the final payoff. In this case, the net incentive value should bias confidence judgments.

In four experiments, we repeatedly found behavioral patterns that confirm the motivational effect of incentives on confidence sensitivity and that pinpoint a biasing effect of incentives in line with our value-confidence interaction hypothesis. We therefore suggest that, similarly to choices and in line with affect-as-information theories (*36*), confidence judgments are biased by incentive-induced affective signals.

RESULTS

We collected data in four experiments in which participants performed different versions of a confidence elicitation task (Fig. 1, table S1, and Materials 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 of gaining or losing a certain payoff (for example, 10¢, 1€, and 2€), which differed between the experiments.

Participants could maximize their chance of gaining (or not losing) the stake by reporting their confidence as accurately and truthfully as possible, because the outcome of the trial was determined by a matching probability (MP) mechanism, a well-validated method from behavioral economics adapted from the Becker-DeGroot-Marschak auction (*37*, *38*). Briefly, the MP mechanism considers participants' confidence reports as bets on the correctness of their answers and implements 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 most truthful confidence estimation (*39*, *40*). The MP mechanism remains incentive-compatible when subjects are not risk-neutral (*40*, *41*). Because this incentivization was implemented after the perceptual choice, it is possible to separately motivate the accuracy of confidence judgments without directly influencing the performance on the perceptual decision. Before the task, participants performed a calibration session, which was used to generate the main task stimuli, such that the subjective difficulties of perceptual choices spanned a predefined range (see Materials and Methods).

Basic features of confidence judgments

As a prerequisite, we assessed the quality of our experimental design and the validity of our experimental variables, irrespective of the effects of monetary incentives. Notably, we show that, in all four experiments, ex ante choice predictions from our psychophysical model closely match participants' actual choice behavior (Supplementary Results). Additionally, we show that in all four experiments, participants' confidence judgments exhibit three fundamental properties (*42*): (i) Confidence ratings correlate with the probability of being correct (this is a natural requirement for the internal consistency of confidence); (ii) the link between confidence ratings and perceptual evidence (see Materials and Methods for definition) is positive for correct and negative for incorrect responses (this follows from the fact that with higher levels of evidence, the probability of individuals being incorrect and very confident in this incorrect response is low); and (iii) the link between evidence and performance differs between high- and low-confidence trials (Supplementary Results). Overall, these preliminary results suggest that the confidence measure elicited in our task actually corresponds to subjects' estimated posterior probability of being correct (*42*, *43*). They also address potential concerns about the validity of confidence elicitation in general (*44*, *45*) and additionally demonstrate that our MP incentivization mechanism did not bias or distort confidence.

Effects of incentives on confidence judgments

Twenty-four subjects participated in our first experiment, where the combination of their choice and confidence ratings could lead, depending on the trial, to a gain or no-gain of 1€, to a loss or no-loss of 1€, or to a neutral outcome (Fig. 1C). To investigate the interaction between incentive motivation and confidence, and compare the predictions of the different theories (Fig. 2A), we implemented linear mixed-effects models, with the net and absolute incentive values as independent variables (see Materials and Methods). In line with our value-confidence interaction hypothesis, our results first show that participants' confidence judgments are specifically modulated by the net incentive value ($\beta_V = 2.06 \pm 0.42$, $P < 0.001$; $\beta_{|V|} = -0.97 \pm 1.03$, $P = 0.38$). Critically, and as expected from our task design, this effect of incentives on confidence is not driven by an effect on performance, given that neither the net value nor the absolute incentive value has any effect on performance ($P > 0.20$ for both).

Effects of incentives on confidence (metacognitive) accuracy

To explore how incentives affect confidence accuracy, we adopted the signal detection theory (SDT) approach developed for metacognition (*7*, *46*, *47*). SDT postulates that both the binary choice and the metacognitive (confidence) estimation are based on the same noisy source of perceptual evidence. The goal of SDT analysis is to estimate from the observed distributions of choices and confidence ratings how this internal signal is used by participants to derive their decisions. Under a few assumptions, the SDT framework can be used to dissociate and measure two components of metacognitive accuracy: the metacognitive bias and the metacognitive sensitivity.

The metacognitive bias is the tendency to give high confidence ratings, all else being equal (*7*). We used, as a measure of this bias, a classical measure of overconfidence (*13*, *41*), computed as the difference between the averaged confidence and the averaged performance. Therefore, a metacognitive bias of zero signals high confidence accuracy, whereas a positive (or negative) calibration signals overconfidence (or underconfidence) and thus lower confidence accuracy.

The metacognitive sensitivity measures the efficacy with which observers' confidence ratings discriminate between their own correct and incorrect answers. We used, as a metric for this sensitivity, the meta-d′, which estimates how much information, in signal-to-noise (d) units, is available for confidence estimation (*46*). Therefore, the higher the meta-d′, the more sensitive an observer's confidence judgment is to the correctness of his or her choice (*7*, *46*, *48*). Our results first show that metacognitive sensitivity (meta-d′) is specifically modulated by the absolute incentive value ($\beta_V = \beta = 0.00 \pm 0.05$, $P = 0.94$; $\beta_{|V|} = 0.34 \pm 0.09$, $P < 0.001$; Fig. 2B); this means that positive and negative incentivization symmetrically improve participants' metacognitive sensitivity compared to the no-incentive condition. We refer to this first effect as the motivational effect of incentives on confidence accuracy. Second, mirroring the effects on confidence judgments, metacognitive bias monotonically increases with the net incentive value ($\beta_V = 2.78 \pm 0.67$, $P < 0.001$; $\beta_{|V|} = -0.71 \pm 1.42$, $P = 0.62$; Fig. 2B). Because participants are overconfident on average, metacognitive bias is thereby improved by loss prospects but paradoxically deteriorated by gain prospects. We refer to this second effect as the biasing effect of incentives on confidence.

Effects of incentives on confidence formation

By assuming that confidence builds on noisy perceptual evidence (*43*, *46*), we expect to observe a positive correlation between confidence and perceptual evidence for correct choices and a negative correlation for incorrect choices [see Sanders *et al.* (*42*) and Fleming and Daw (*43*) and Supplementary Results]. Another way of investigating the consequences of confidence incentivization on metacognitive accuracy is to assess how incentives modulate the relationship between confidence and evidence for correct and incorrect answers: Incentive effects can affect confidence per se (suggesting a simple bias of confidence) or influence the relationship between confidence and evidence (suggesting that incentives affect the integration of evidence in the formation of the confidence signal). Although similar in essence to the metacognitive metrics (bias and sensitivity) used above, this approach is model-free and does not rely on some of the assumptions required for the meta-d′ (*7*, *46*). Thus, for each individual and each incentive level, we built a multiple linear regression modeling confidence ratings as a combination of a confidence baseline and two terms capturing the linear integration of perceptual evidence for correct and incorrect answers (see Materials and Methods). Regression coefficients were estimated at the individual level for each incentive level, and the effect of incentives on

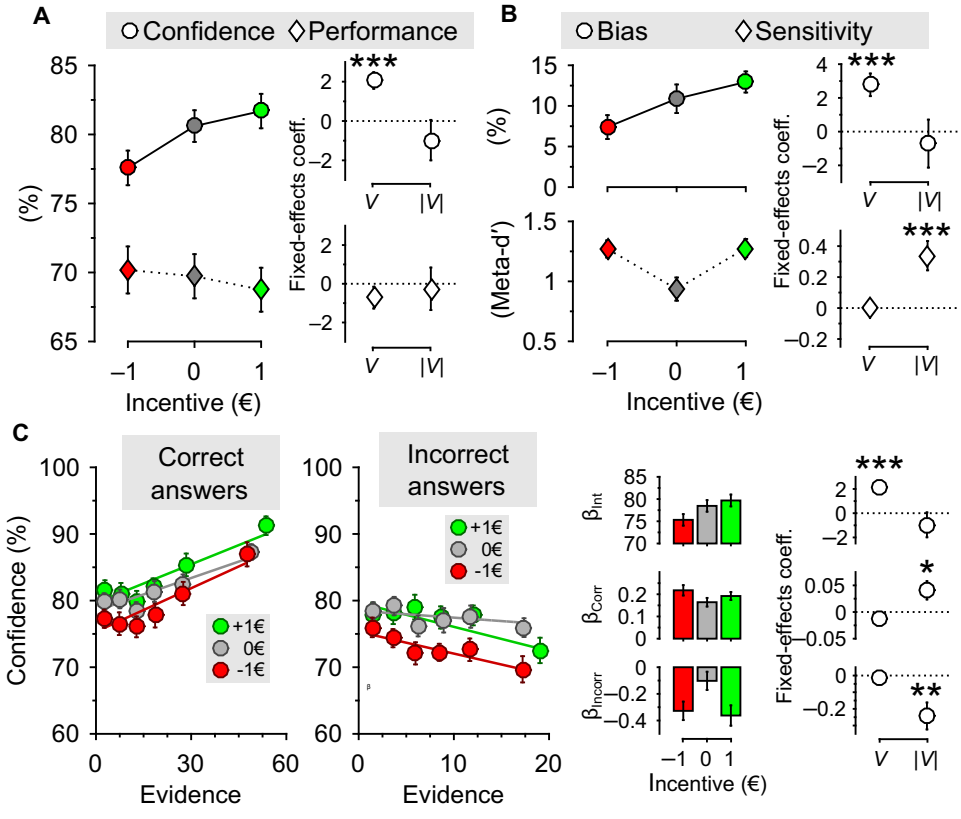

Fig. 2. Experiment 1. (**A**) Incentive effects on behavior. Reported confidence (dots) and performance (diamonds)—that is, % correct—as a function of incentives. (**B**) Incentive effects on confidence (metacognitive) accuracy. Computed bias (top, dots) and meta-d′ (diamonds) as a function of incentives. The insets presented on the right-hand side of the graphs (A and B) depict the results of the linear mixed-effects model, estimated for each behavioral (A, top: confidence; bottom: performance) and metacognitive (B, top: bias; bottom: sensitivity) measure. (**C**) Incentive effects on confidence formation. Linking incentives, evidence, and confidence for correct (left) and incorrect (right) answers. In those two panels, the scatterplots display reported confidence as a function of evidence for the different incentive levels. The solid line represents the best linear regression fit at the population level. The histograms represent the intercepts (top) and slope for correct (middle) and incorrect (bottom) answers of this relationship, estimated at the individual level and averaged at the population level. The insets presented on the right-hand side of the graph depict the results of the linear mixed-effects model, estimated for each parameter of this regression, that is, intercept (top) and slope for correct answers (middle) and for incorrect answers (bottom). *V*, net incentive value; |*V*|, absolute incentive value. Error bars indicate intersubject SEM. **P* < 0.05, ***P* < 0.01, ****P* < 0.001.

the different regression coefficients was subsequently tested in our linear mixed-effects model. Our results show a clear dissociation between the motivational and biasing effects of incentives on confidence formation. On the one hand, the absolute incentive value affects the slopes of those regressions: In both cases, gains and losses increase the linear relationship between confidence and evidence compared to no incentives (correct answers: $\beta_{|V|} = 0.04 \pm 0.02$, $P < 0.05$; incorrect answers: $\beta_{|V|} = -0.24 \pm 0.08$, $P < 0.01$; Fig. 2C). On the other hand, the net incentive value affects the intercept of those regressions ($\beta_V = 2.18 \pm 0.47$, $P < 0.001$; Fig. 2C). This indicates that while the motivational effect of incentives actually influences the way confidence is built from evidence by increasing the weight of evidence in the ratings in the opposite direction for correct and incorrect answers, the biasing effect of incentives appears to be a purely additive effect of incentives on confidence, unrelated to the amount of evidence. These results therefore confirm and extend the reported biasing effects of incentives on overconfidence (metacognitive bias) and the motivational effects of incentives on metacognitive sensitivity. To further investigate how incentives influence confidence, and to control for alternative explanations, we next conducted three additional experiments.

The effects of incentives without incentivizing confidence judgments

To rule out that participants deliberately and strategically increase their confidence with net incentive value, due to some misconceptions induced by the incentivization (that is, MP) mechanisms, we collected data from 21 new participants who performed a second task without MP incentivization (that is, a performance task) in addition to our standard confidence task. In the performance task, confidence is simply elicited with ratings after choice; confidence accuracy is not incentivized with the MP mechanisms, and subjects are only rewarded according to their choice performance—correct/incorrect (see Materials and Methods and Fig. 3A). Still, in line with the value-confidence hypothesis, confidence is found to be specifically modulated by the net incentive value in both the confidence and the performance task (confidence task: $\beta_V = 2.47 \pm 0.68$, $P < 0.001$; performance task: $\beta_V = 1.88 \pm 0.59$, $P < 0.01$; Fig. 3B), while incentives have no effect on performance in either task (as expected from the task design; $P > 0.38$ for all).

Regarding confidence accuracy, the effect of the net incentive value affects metacognitive bias in both tasks, but merely as a trend (confidence task: $\beta_V = 2.01 \pm 1.24$, $P = 0.11$; performance task: $\beta_V = 1.75 \pm 0.89$,

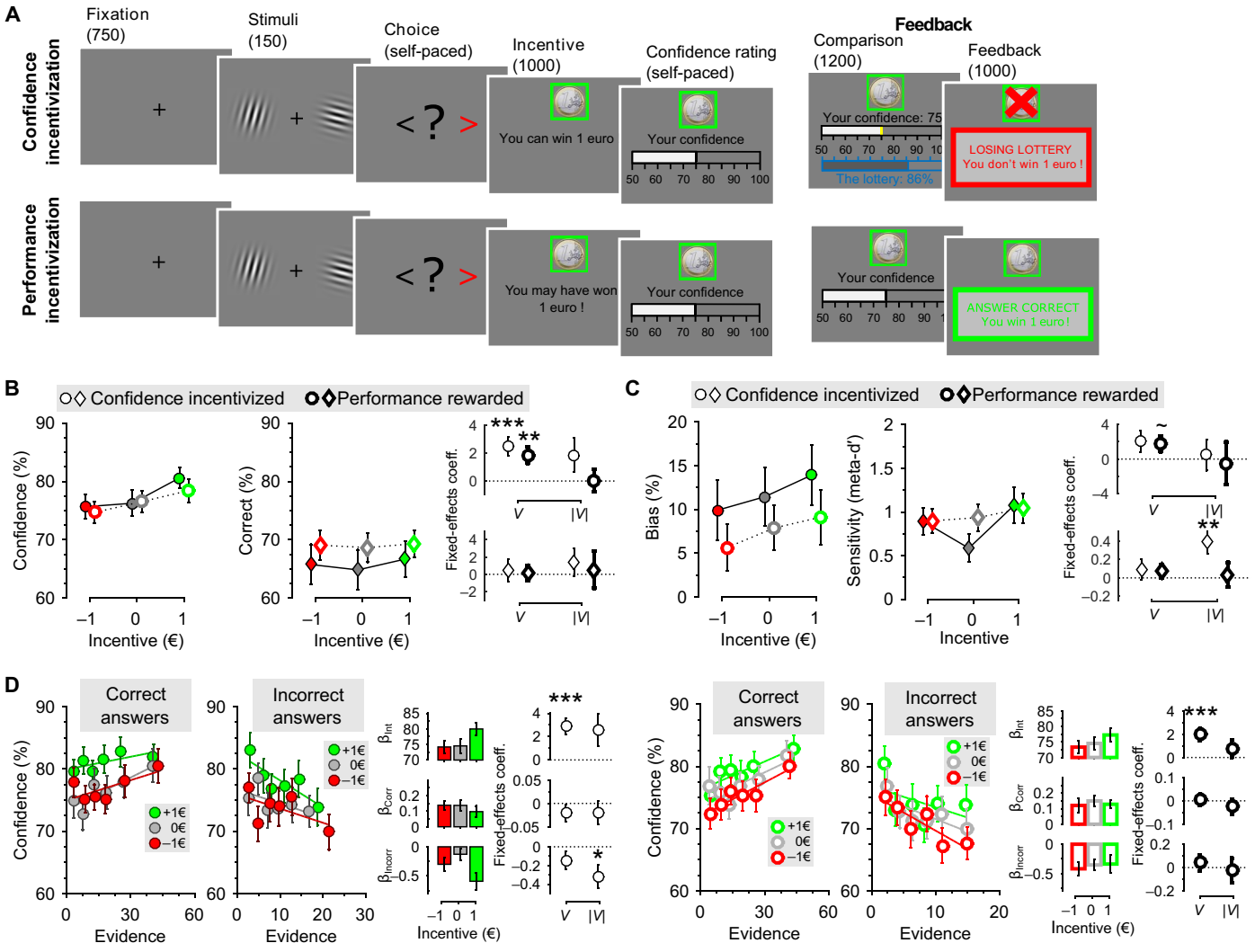

Fig. 3. Experiment 2. (**A**) In experiment 2, participants performed two versions of the task: one in which confidence reports were incentivized (top row), and one in which the performance (that is, correctness of the binary choice) was incentivized (bottom row). See also table S1. Successive screens displayed in one trial are shown from left to right, with durations in milliseconds. Participants viewed a couple of Gabor patches displayed on both sides of a computer screen, and judged which had the highest contrast. They were then presented with a monetary stake (in a green frame for gain, gray for neutral, and red for losses) and asked to report their confidence *C* in their answer on a scale from 50 to 100%. (**B**) Incentive effects on behavior. Reported confidence (dots, leftmost scatterplot) and performance (diamonds, rightmost scatterplot)—that is % correct—as a function of incentives. (**C**) Incentive effects on confidence accuracy. Computed metacognitive bias (dots, leftmost scatterplots) and sensitivity (diamonds, rightmost scatterplots) as a function of incentives. The insets presented on the right-hand side of the graphs (A and B) depict the results of the linear mixed-effects model, estimated for each behavioral (A, top: confidence; bottom: performance) and metacognitive (B, top: bias; bottom: sensitivity) measure. Empty markers with thick edges indicate the performance rewarded task. (**D**) Incentive effects on confidence formation. Linking incentives, evidence, and confidence for the confidence incentivized (left half) and the performance rewarded (right half) tasks, for both correct (left scatterplot) and incorrect (right scatterplot) answers. In those two panels, the scatterplots display reported confidence as a function of evidence for the different incentive levels. The solid line represents the best linear regression fit at the population level. The histograms represent the intercepts (top) and slope for correct (middle) and incorrect answers (bottom) of this relationship, estimated at the individual level and averaged at the population level. The insets presented on the right-hand side of the graph depict the results of the linear mixed-effects model, estimated for each parameter of this regression, that is, intercept (top) and slope for correct answers (middle) and for incorrect answers (bottom). *V*, net incentive value; |*V*|, absolute incentive value. Error bars indicate intersubject SEM. ˜*P* < 0.10; **P* < 0.05; ***P* < 0.01; ****P* < 0.001.

P = 0.05; Fig. 3C). The motivational effect of the absolute incentive value on metacognitive sensitivity is replicated when confidence is incentivized, but not when performance is incentivized (confidence task: $\beta_{|V|} = 0.40 \pm 0.14$, $P < 0.01$; performance task: $\beta_{|V|} = 0.04 \pm 0.13$, $P = 0.79$; Fig. 3C). We then replicate and extend the findings of the first experiment on the confidence formation model (Fig. 3D): When biasing effects of the net incentive value are present (in both tasks), they affect the intercept of the confidence formation model (both *P*'s < 0.001). On the other hand, the motivational effects of the absolute incentive value are only found on the slope of incorrect trials in the confidence task (incorrect answers; confidence task: $\beta_{|V|} = -0.317 \pm 0.13$, $P < 0.05$; performance task: $\beta_{|V|} = -0.03 \pm 0.10$, $P = 0.81$). Again, this means that the motivational effect of the incentivization of confidence accuracy is underpinned by a better integration of perceptual evidence in the confidence rating when stakes

increase, whereas this effect is absent in the task where confidence accuracy is not incentivized. In sum, these results indicate that the biasing effects of incentives on confidence judgments are not induced by the incentivization mechanism and that the motivational effects of incentives on confidence and metacognitive accuracy are only found when confidence is incentivized.

Dissociating incentive value effects from simple valence effects

To demonstrate that the motivational and biasing effects of incentives are due to incentive values, rather than to simple valence (gain/loss) effects, we next invited 35 subjects to participate in a third experiment, where incentives for confidence accuracy varied in both valence (gains and losses) and magnitude (1€ versus 10¢) (see table S1). We modified our linear mixed-effects models to include a valence variable (=1 if incentives are positive and 0 if negative, indexed by +/−), in addition to the net and absolute incentive value. Results show that both the net incentive value and the valence variable affect confidence judgments (β_V = 1.02 ± 0.38, P < 0.01; $\beta_{+/-}$ = 4.01 ± 1.11, P < 0.001; Fig. 4A). This means that the biasing effects of incentives previously reported are not simply due to an effect of valence but are truly underpinned by the net incentive value. Again, as designed and expected, no effect of incentives is found on performance (all P's > 0.22). The linear effect of the net incentive value transfers to the metacognitive bias (β_V = 2.15 ± 0.97, P < 0.05; Fig. 4B). Note that we do not find significant effects of the absolute value of incentives on metacognitive sensitivity (all P's > 0.25; Fig. 4B). This difference with the results of the two previous experiments can be explained by the lack of a neutral incentive condition in the present experiments. This means that motivational effects previously reported would be primarily due to the mere presence of incentives.

Replicating our previous finding, the biasing effect of the net incentive value is found to be independent from the amount of evidence, affecting the intercepts of the linear relationship between evidence and confidence (intercept: β_V = 1.34 ± 0.50, P < 0.01; Fig. 4C). No effect of incentives is found on the slopes characterizing the integration of evidence in confidence judgments (all P's > 0.11). In sum, the results from this third experiment replicate the biasing effect of the net incentive value on confidence and further demonstrate that these effects depend on the magnitude of incentives.

Accounting for difference between gain and loss in effect on confidence

While the effect of the net incentive value on confidence and metacognitive bias revealed in our first three experiments appeared robust and

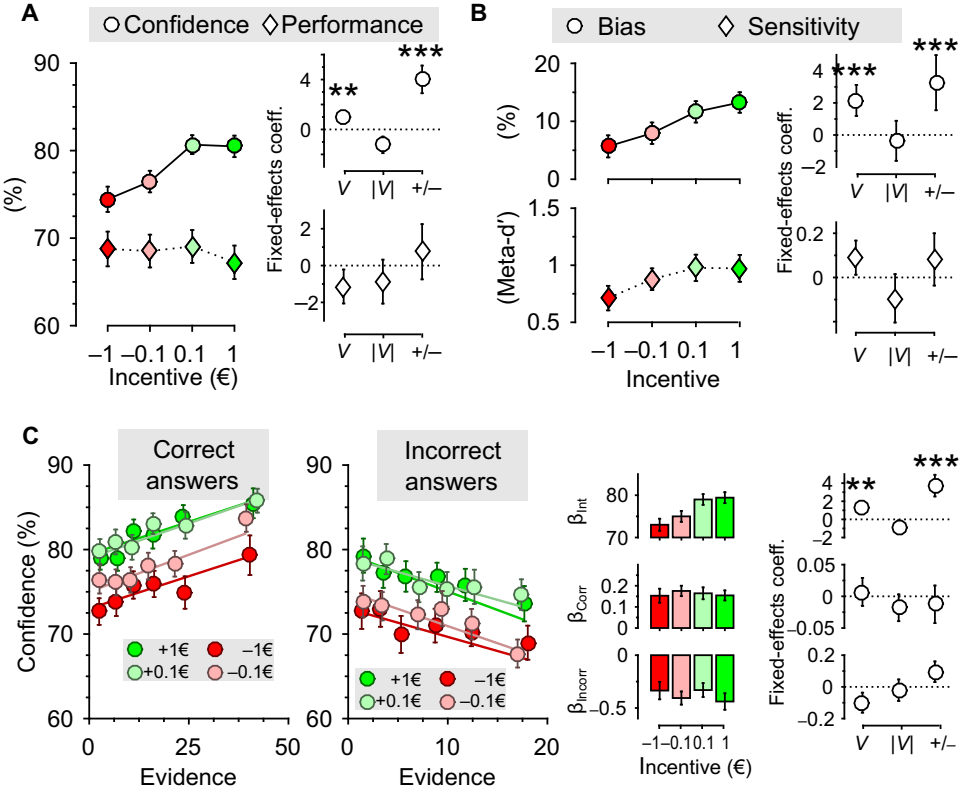

Fig. 4. Experiment 3. (**A**) Incentive effects on behavior. Reported confidence (dots) and performance (diamonds)—that is, % correct—as a function of incentives. (**B**) Incentive effects on confidence (metacognitive) accuracy. Computed bias (top, dots) and meta-d′ (diamonds) as a function of incentives. The insets presented on the right-hand side of the graphs (A and B) depict the results of the linear mixed-effects model, estimated for each behavioral (A, top: confidence; bottom: performance) and metacognitive (B, top: bias; bottom: sensitivity) measure. (**C**) Incentive effects on confidence formation. Linking incentives, evidence, and confidence for correct (left) and incorrect (right) answers. In those two panels, the scatterplots display reported confidence as a function of evidence for the different incentive levels. The solid line represents the best linear regression fit at the population level. The histograms represent the intercepts (top) and slope for correct (middle) and incorrect answers (bottom) of this relationship, estimated at the individual level and averaged at the population level. The insets presented on the right-hand side of the graph depict the results of the linear mixed-effects model, estimated for each parameter of this regression, that is, intercept (top) and slope for correct answers (middle) and for incorrect answers (bottom). *V*, net incentive value; |*V*|, absolute incentive value; +/–, incentive valence. Error bars indicate intersubject SEM. ***P* < 0.01; ****P* < 0.001.

replicable, it seemed to be driven by the loss frame. This could mean that this biasing effect is purely restricted to the loss frame. However, an alternative hypothesis is that subjects are simply less sensitive to gains, as suggested by prospect theory (*49*). To distinguish between those two hypotheses, we invited 24 subjects to participate in a final study, which included higher stakes (10¢, 1€, 2€) in both gain and loss frames (table S1). In this case, our linear mixed-effects model included three independent variables: two variables accounted for the signed incentive magnitude in the gain frame (*V*+) and in the loss frame (*V*−); in addition, and in line with the previous experiment, the third variable captured the effect of the valence framing (+/−). Our results reveal a significant effect of the incentive magnitude on confidence, in both the gain and loss frames (β_{V+} = 0.79 ± 0.25, $P < 0.001$; β_{V-} = 1.22 ± 0.38, $P < 0.001$; Fig. 5A), and no effects of incentives on performance (all *P*'s > 0.45). This result confirms our initial hypothesis: Following expected values, higher incentives seem to bias confidence judgments upward in a gain frame and downward in a loss frame. Yet, it is worth noting that the absolute effect size is about 50% larger in the loss domain than in the gain domain. This is consistent with the idea of loss aversion: People prefer avoiding losses to acquiring equivalent gains; hence, loss prospects have stronger motivational values than equivalent gain prospects (*49*).

Similar to our third experiment, no motivational effect of incentives is detectable on metacognitive sensitivity (all *P*'s > 0.11; Fig. 5B), suggesting that it is mostly driven by the incentive versus no-incentive contrast. Slightly departing from what we observed in the previous experiments, the effects of incentives on metacognitive bias are, this time, mostly driven by the valence variable ($\beta_{+/-}$ = 4.26 ± 1.72, $P < 0.05$; Fig. 5B). Given that this measure combines the confidence and performance variance, and that the presence of six incentive levels decreases the number of trials used to estimate it, we interpret the absence of an incentive magnitude effect on metacognitive bias as a lack of power. Supporting this interpretation, the biasing effects can be found on the intercept of our confidence formation model, a more sensitive measure of our bias (β_{V+} = 0.72 ± 0.40, $P = 0.08$; β_{V-} = 1.26 ± 0.58, $P < 0.05$; Fig. 5C). This last set of results replicates, for the fourth time, the biasing effects of incentives on confidence and confirms that both monetary gains and losses contribute to biasing confidence in perceptual decisions.

Estimating the costs of confidence biases

To investigate the consequences of the incentive bias on confidence that we demonstrated in this report, we derived the expected costs of the interaction between confidence and incentives (see Materials and

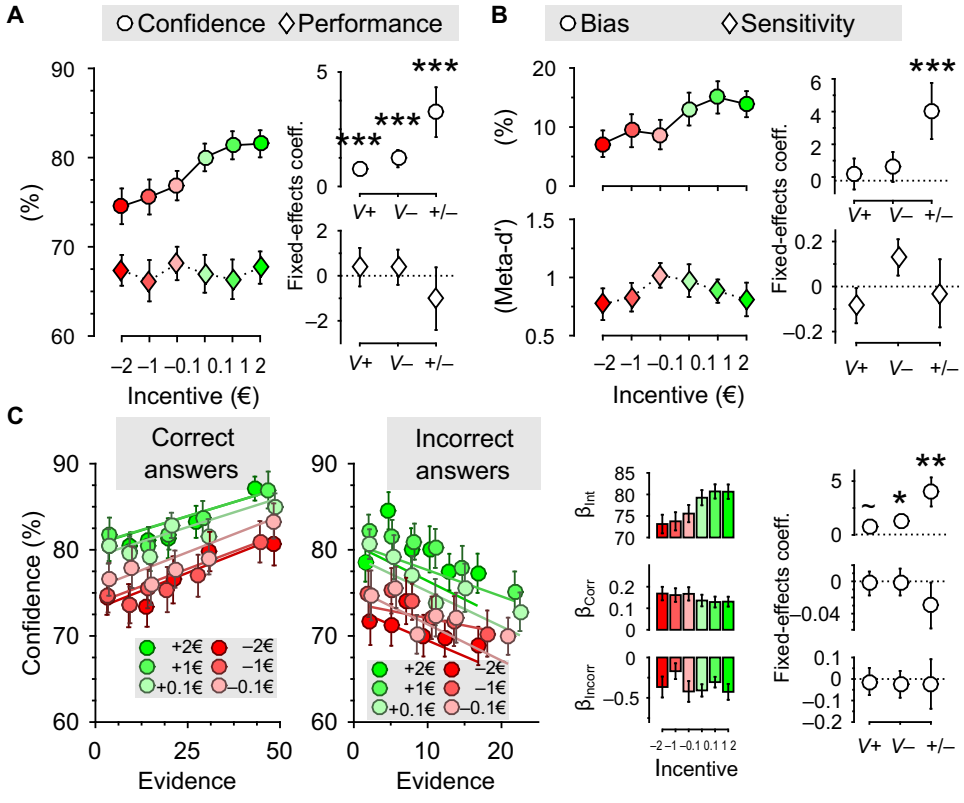

Fig. 5. Experiment 4. (**A**) Incentive effects on behavior. Reported confidence (dots) and performance (diamonds)—that is, % correct—as a function of incentives. (**B**) Incentive effects on confidence (metacognitive) accuracy. Computed bias (top, dots) and meta-d' (diamonds) as a function of incentives. The insets presented on the right-hand side of the graphs (A and B) depict the results of the linear mixed-effects model, estimated for each behavioral (A, top: confidence; bottom: performance) and metacognitive (B, top: bias; bottom: sensitivity) measure. (**C**) Incentive effects on confidence formation, linking incentives, evidence, and confidence for correct (left) and incorrect (right) answers. In those two panels, the scatterplots display reported confidence as a function of evidence for the different incentive levels. The solid line represents the best linear regression fit at the population level. The histograms represent the intercepts (top) and slope for correct (middle) and incorrect answers (bottom) of this relationship, estimated at the individual level and averaged at the population level. The insets presented on the right-hand side of the graph depict the results of the linear mixed-effects model, estimated for each parameter of this regression, that is, intercept (top) and slope for correct answers (middle) and for incorrect answers (bottom). *V*+, net incentive value for gains; *V*–, net incentive value for losses; +/–, incentive valence. Error bars indicate intersubject SEM. ~*P* < 0.10, **P* < 0.05, ***P* < 0.01, ****P* < 0.001.

Methods). In the current setting, and with the effect size observed in experiment 1 ($\beta_V = 2.78$), this bias would have modest consequences for the payoffs of well-calibrated participants (a loss of roughly 0.1% winning probability for an incentive of 1€ compared to the optimal policy). However, the derivations also show that the consequences of this bias can be more severe in the presence of an existing bias such as overconfidence, because the costs of biases are multiplicative rather than additive (see Materials and Methods and Fig. 6B). Together with the overconfidence observed in the absence of incentives in experiment 1 (11%), an incentive of 1€ causes an additional 0.75% decrease in winning probability, resulting in a total cost of 2% decreased winning probability. We additionally assessed the total financial cost caused by the combination of overconfidence and incentive bias (Fig. 6C). These results illustrate how incentivizing confidence with gains (or losses) decreases the financial losses induced by underconfidence (or overconfidence) but concurrently increases financial losses induced by overconfidence (or underconfidence).

DISCUSSION

Here, we combined a perceptual decision task and an auction procedure inspired by behavioral economics (*37*, *38*) to investigate how monetary incentives influence confidence. In addition to replicating important statistical features common to most of the dominant models of confidence formation (*43*), we reveal and dissociate two effects of monetary incentives on confidence accuracy.

The first effect is a motivational effect of incentives: In line with theories of rational decision-making and motivation, incentivizing confidence judgments improves metacognitive sensitivity. This means that high (or low) confidence is more closely associated with correct (or incorrect) decisions when confidence reports are incentivized, regardless of the valence or magnitude of the incentive. This extends a recent study reporting a similar effect of incentivization on discrimination (a measure closely related to sensitivity, assessing how confidence discriminates between correct and incorrect answers), but limited to the gain domain (*41*). This also confirms that the MP mechanism is particularly well suited to investigation of confidence incentivization (*41*, *47*). Here, we further show that this motivational effect of incentives is underpinned by a better integration of perceptual evidence in confidence judgments when stakes increase. Although these motivational effects were clear in experiments 1 and 2, where incentivized conditions (1€ gain or loss) were compared to a non-incentivized condition, they did not extend to experiments 3 and 4, where different levels of incentives were compared. This discrepancy could be explained either by a lack of power to detect these effects as a result of fewer trials per incentive condition or by psychological effects related to higher incentive magnitudes [for example, the participants could choke under pressure (*50*)]. Note that potentially detrimental effects of high incentives on metacognitive performance have been reported in the domain of perceptual awareness (*51*).

The second effect, the biasing effect of incentives, is more striking: Confidence judgments are parametrically biased by the net incentive value. The prospect of gains increases confidence, while the prospect of losses decreases confidence. Because people generally exhibited overconfidence in our experiment, gain prospects detrimentally increased the overconfidence bias, while prospects of losses reduced this bias and improved confidence accuracy. There are two possible interpretations for the effects in the loss frame: (i) loss prospects can truly improve calibration, or (ii) symmetrically to the gain condition, they simply bias confidence downward, which happens to correct overconfidence. Although the data presented here cannot tease apart those two hypotheses, further research, for example, translating the current design in a context where individuals are underconfident, could straightforwardly address this question. As opposed to the motivational effect, the biasing effect of incentives was purely additive, that is, independent of the amount of evidence on which decisions and confidence judgments are based. The biasing effect was also found to be incidental, that is, also present when performance, but not confidence, was incentivized. We show that this bias is unpredicted by motivated cognition theories such as the desirability bias (*26*), which predicts that the overconfidence bias would also increase with negative incentive values, because avoiding a loss is desirable. This biasing effect is also unpredicted by the theories of rational decision-making and motivation, which predict decreased overconfidence with increased positive incentive values because it would lead to a higher reward (as incentivized by the MP mechanism). Yet, the biasing effect of incentives is in line with the value-confidence hypothesis. One plausible interpretation for this effect is an affect-as-information effect: People use their momentary affective states as information in decision-making (*36*), which, in our case, means that they integrate the trial expected value into their confidence judgment.

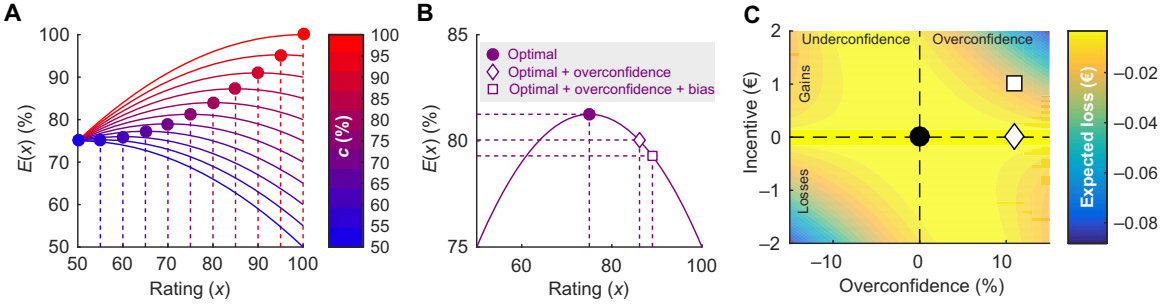

Fig. 6. Incentive bias costs. (**A**) Expected probability of winning $E(x)$ induced by the MP mechanism, as a function of the chosen rating x for several levels of underlying confidence c. The dots indicate the highest point of each curve, which correspond to $x = c$. (**B**) Effects of a combination of biases (11% overconfidence + 3% extra bias due, for example, to an incentive of 1€) on the expected probability of winning $E(x)$. The values of the overconfidence and incentive biases correspond to the values observed in experiment 1. (**C**) Expected loss induced by a combination of an overconfidence bias (x axis) and an additional incentive bias ($\beta_V = 3\%$) for several levels of incentives (y axis). Note that incentivizing confidence with gains (or losses) decreases the losses induced by underconfidence (or overconfidence) (top left and bottom right corners). On the contrary, incentivizing confidence with gains (or losses) increases losses induced by overconfidence (or underconfidence) (top left and bottom right corners). The markers (dot, diamond, and square) correspond to the parameters used in (B).

These results and interpretations fit with recent reports showing that negative affective states (such as worry) decrease overconfidence (*28*), while positive affective states (such as joy) increase overconfidence (*27*). The reported effects of incentives on confidence also confirm that confidence judgments not only represent rational estimates of the probability of being correct (*3*) but also integrate information and potential biases processed after a decision is made (*43*, *52*). These results therefore provide additional evidence in favor of second-order models of confidence, which propose that confidence builds on samples of evidence different from the ones used to render the decision (*43*).

To incentivize confidence reports, we used a mechanism inspired by Becker-DeGroot-Marschak auction procedures (*37*, *38*), referred to as reservation probability or MP, which conveniently allowed us to manipulate the monetary stakes on a trial-by-trial basis. Contrary to other incentivization methods such as the quadratic scoring rule (QSR), the MP mechanism is valid under simple utility maximization assumptions, that is, remains incentive-compatible when subjects are not risk-neutral (*40*, *41*). The MP mechanism is even incentive-compatible when considering probability distortions, on the assumption that both subjective (confidence) and objective (lotteries) probabilities are transformed identically (*53*, *54*). This implies that the incentive bias on confidence uncovered in this study cannot be attributed to factors such as asymmetries in risk attitude between gain and loss frames (*55*). Yet, in ecological situations, this bias could easily be worsened or corrected by effects of risk attitude on confidence (*56*).

Several studies have investigated the impact of different incentivization mechanisms on subjective probability judgments (confidence or belief) and report that MP is among the best methods available, at both the theoretical and experimental levels (*40*, *41*, *53*), and is particularly well suited for SDT analyses (*47*). MP is truly incentive-compatible and elicits an unbiased estimator of confidence in the absence of any bias induced by monetary incentives. However, the presence of such a bias, as demonstrated in the present report, challenges the ability of this mechanism to elicit unbiased confidence judgments.

In this collection of experiments, we only used relatively small monetary amounts as incentives; how the motivational and biasing effects of incentives scale when monetary stakes increase remains an open question. Critically, higher stakes may also affect physiological arousal, which can influence confidence and interoceptive abilities (*30*, *57*). In general, the effects of incentives on confidence accuracy could also be mediated by interindividual differences in metacognitive or interoceptive abilities (*57*, *58*) and by incentive motivation sensitivity (*59*). Because our subject sample was mostly composed of university students, the generalization of those findings to a wider population will have to be assessed in further studies.

The mere notion of confidence biases, notably overconfidence, and the actual conditions under which they can be observed sparked an intense debate in psychophysics (*14*, *60*, *61*) and evolutionary theories (*62*, *63*). Critically, here, confidence accuracy was properly incentivized; hence, deviations from perfect calibration can be appropriately interpreted as cognitive biases (*63*). The striking effects of net incentive values on confidence seem to make sense when considering an evolutionary perspective: In natural settings, whereas overconfidence might pay off when prospects are potential gains [for example, when claiming resources (*62*)], a better calibration might be more appropriate when facing prospects of losses (for example, death or severe injuries), given their potential dramatic consequences on reproductive chances. The observed valence difference in the effect of incentive magnitude—higher in the loss domain than in the gain domain—seems to mimic valence asymmetries observed in economic decision-making theories such as prospect theory (*49*).

How confidence is formed in the human brain and how neurophysiological constraints explain biases in confidence judgments remain an open question (*3*, *64*). Although functional and structural neuroimaging studies initially linked confidence and metacognitive abilities to dorsal prefrontal regions (*4*), confidence activations were also recently reported in the ventro-medial prefrontal cortex (*31*, *32*) and in striatal and mesolimbic regions (*33*, *34*). This network has been consistently involved in motivation and value-based decision-making (*35*). It is therefore possible that this network plays a role in the motivational and biasing effects of incentives on confidence. 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 into the cause of confidence miscalibration in healthy and pathological settings. For instance, high monetary incentives in financial or managerial domains may create or exaggerate overconfidence, leading to overly risky and suboptimal decisions. In the clinical context, inflated levels of overconfidence in pathological gamblers (*65*) could be amplified by high monetary incentives, contributing to compulsive gambling in the face of great loss. Moreover, if value-induced affective states modulate confidence judgments, then other disorders with abnormal incentive processing such as addictions, mood disorders, obsessive-compulsive disorder, and schizophrenia could be at particular risk for confidence miscalibration (*66*–*68*). Field experiments and clinical research will be needed to further explore the individual and societal consequences of the interactions between incentive motivation and confidence accuracy.

MATERIALS AND METHODS

Subjects

All studies were approved by the local ethics committee of the University of Amsterdam Psychology Department. All subjects gave informed consent before partaking in the study. The subjects were recruited from the laboratory's participant database (www.lab.uva.nl). A total of 104 subjects took part in this study (see table S1). 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 (www.vislab.ucl.ac.uk/cogent.php). In all four experiments, trials of the confidence incentivization task shared the same basic steps (Fig. 1A): 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 then presented with a monetary stake (1000 ms, accompanied by the sentence "You can win[/lose] X euros") 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 initial position of the cursor was randomized between 65 and 85% to avoid anchoring of answers on 75%. 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 into 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, feedback was displayed, which depended on the correctness of the initial choice. Back at the resolution step, if L happened to be higher than C, the lottery was implemented. A wheel of fortune, with an L% chance of winning, was displayed and played; the lottery arm spun 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, feedback informed the subject whether they had won or lost the lottery.

Subjects 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. Because of the MP procedure, the strategy to maximize one's earnings is to always report one's subjective probability of being correct as truthfully and accurately as possible on the confidence scale (Supplementary Materials).

Subjects were explicitly informed of this. In addition to extensive instructions explaining the MP procedure, participants gained direct experience with this procedure through a series of 24 training trials that did not count toward final payment.

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 (1200 ms). Additionally, 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. There was no difference in our participants' behavior when the extended or short version of our task was used.

In the performance version, the MP mechanism was omitted, but the layout was similar to the short version (see Fig. 3A). The monetary stake screen was accompanied by a different sentence (You may have won[/lost] X euros). The lottery draw/comparison step was replaced with a screen of similar duration (1200 ms), simply displaying the confidence scale and the chosen rating. A feedback screen displayed the correctness of the answer and the trial outcome at every trial (1000 ms).

Stimuli and design

Participants initially performed a 144-trial calibration session (~5 min), where they only performed the Gabor contrast discrimination task, without an incentive or confidence measure (Fig. 1A). During this calibration, the distribution of contrast difference (that is, difficulty) was adapted every 12 trials following a staircase procedure (see the Supplementary Materials) such that performance reached approximately 70% correct.

The calibration data were used to estimate individual psychometric function

$$p(\mathrm{ch_L}) = 1 + \exp(-\mu - \sigma \times (C_\mathrm{L} - C_\mathrm{R}))^{-1}$$

where $p(\mathrm{ch_L})$ is the probability of subjects choosing the left Gabor, and C_L and 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 [that is, known $p(\mathrm{ch_L})$; see table S1] for all incentive levels. After the first session of the confidence task, μ and σ were reestimated for each session from the data of the preceding session (experiments 1, 2, and 4) or from a new calibration session (experiment 3).

Optimal confidence rating in an MP elicitation mechanism

Here, we provide a simple and accessible version of the demonstration of the incentive compatibility of the MP mechanism.

Let x be potential ratings, l be the random lottery, and c be the true probability of being correct. The random lottery is drawn from the uniform distribution in the interval [0.5 1].

The MP incentivization mechanism considers two mutually exclusive scenarios: The random lottery l is smaller or bigger than the reported rating x. The probabilities associated with these two events are as follows:

$p(l < x)$, that is, the probability that the random lottery l is smaller than the rating x can be written as

$$p(l < x) = \frac{x - 0.5}{0.5} = 2x - 1 \quad (1)$$

$p(l > x)$, that is, the probability that the random lottery l is bigger than the rating x can be written as

$$p(l > x) = \frac{1 - x}{0.5} = 2 - 2x \quad (2)$$

Now, the expected probability of winning is the sum of two terms: $p(\mathrm{w_A})$, the probability of winning as a result of a correct answer, and $p(\mathrm{w_L})$, the probability of winning as a result of the random lottery.

$$E(x) = p(\mathrm{w_A}) + p(\mathrm{w_L}) \quad (3)$$

The first term is basically the multiplication of the probability that the random lottery is smaller than the rating (for the answer to determine the gain) by the probability that the answer is correct (c).

$$p(\mathrm{w_A}) = p(l < x) \times c = (2x - 1)c \quad (4)$$

The second term (Eq. 3) is basically the multiplication of the probability that the random lottery is bigger than the rating (for the lottery to determine the gain), by the expected value of the lottery $E(l|x, c)$.

$$E(l|x, c) = \int_x^1 \frac{t}{1 - x}\, dt = \frac{1 + x}{2} \quad (5)$$

Hence

$$p(\mathrm{w_L}) = p(l > x) \times E(l|x, c) = (2 - 2x)\frac{1 + x}{2} \quad (6)$$

Combining Eqs. 3, 4, and 6, we get

$$\begin{aligned} E(x) &= (2x - 1)c + (2 - 2x)\frac{1 + x}{2} \\ &= (2x - 1)c + (1 - x)(1 + x) = -x^2 + 2xc + 1 - c \end{aligned} \quad (7)$$

Therefore, $E(x)$ is an inverse quadratic function, whose only maximum x_{MAX} is such that

$$E'(x_{MAX}) = 0 \quad (8)$$

Simply computing the derivative of E using Eq. 7, we have

$$E'(x_{MAX}) = -2x_{MAX} + 2c \quad (9)$$

Finally, Eqs. 8 and 9 imply

$$x_{MAX} = c \quad (10)$$

Therefore, to maximize the probability of winning $E(x)$ (that is, maximize the expected outcome), the best possible rating is equal to the true probability of being correct $x = c$ (that is, the unbiased confidence). This proves the incentive compatibility of the confidence elicitation mechanism. Figure 6A depicts the expected probability of winning $E(x)$ as a function of the chosen rating x for several levels of underlying confidence c.

Intuitively, if subjects report $x > c$ (that is, report higher confidence than they actually truly experience), they potentially miss all lotteries defined by $c < l < x$, which would actually give them a higher objective probability of winning the monetary stake than their true confidence c. Likewise, if subjects report $x < c$ (that is, report lower confidence than they actually truly experience), they may face all lotteries l defined by $x < l < c$, which would give them a lower objective probability of winning the monetary stake than their true confidence c. Therefore, to get the highest possible payout, subjects should truthfully report their best estimate of their subjective probability of being correct, that is, their confidence $x = c$.

Metacognitive metrics

We used two components of metacognition: metacognitive bias and metacognitive sensitivity. Metacognitive bias is obtained by computing the difference between the mean confidence and the mean accuracy.

$$\text{Bias} = \frac{1}{n}\sum_{k=1}^{n} C_k - \frac{1}{n}\sum_{k=1}^{n} P_k$$

where n is the total number of trials, C_k is the reported confidence at trial k, and P_k is the performance at trial k (1 for a correct answer and 0 for an incorrect answer).

Metacognitive sensitivity was measured as the meta-d′, a new metric introduced by Maniscalco and Lau (*46*). Meta-d′ defines the level of d′ that an SDT ideal observer would need to generate an observed set of confidence ratings, given an observed set of choices. Meta-d′ was computed using the MATLAB code of Maniscalco and Lau (*46*) available on their website (www.columbia.edu/~bsm2105/type2sdt/). Critically, as opposed to most other metrics of confidence accuracy, the meta-d′ is not influenced by the response bias (such as average confidence level) (*7*, *46*, *48*). Metacognitive efficiency (computed as meta-d′/d′) is often used to assess the relative efficiency of metacognition with respect to performance. Yet, as expected from our task design (that is, the incentives being uncovered after the binary choice), the binary choice performance (that is, the ability to distinguish stimuli, quantified by d′) is not affected by the incentive level in any of the tasks (see Materials and Methods and Supplementary Results), so we chose to run our analyses with the meta-d′ as a measure of metacognitive accuracy.

However, to provide additional evidence that our results were not due to any effects of incentive on first-level performance, all analyses with meta-d′ as the independent variable were replicated using metacognitive efficiency (that is, meta-d′/d′) as an alternative independent variable.

Finally, note that all results obtained with meta-d′ were also replicated with a very simple (but not bias-free) metric of sensitivity, computed as the difference between the average confidence for correct answers and the average confidence for incorrect answers.

Linking confidence and perceptual evidence

Following previous studies (*42*), we computed the perceptual evidence by normalizing the unsigned difference of the two Gabors' contrast intensity by their sum to adjust for saturation effects.

$$\text{Evidence} = 100 \times \left|\frac{G_R - G_L}{G_R + G_L}\right|$$

where G_S is the contrast intensity of the Gabor displayed on side S (S = L for left and S = R for right) of the screen. For each individual, and each incentive level, confidence was then regressed against this measure of perceptual evidence for both correct and incorrect choices using the following regression model

$$\text{Confidence} = \beta_{int} + \beta_{Corr} \times I_{Corr} \times \text{evidence} + \beta_{Incorr} \times I_{Incorr} \times \text{evidence}$$

where I_{Corr} and I_{Incorr} are indicative (that is, dummy) variables for correct and incorrect binary perceptual decisions. The parameters (β_{int}, β_{Corr}, and β_{Incorr}) were estimated for each individual and incentive level and then fed to linear mixed-effects models (see the next paragraph) to test the influence of incentive levels on confidence bias (or intercept, β_{int}) and on how confidence integrates perceptual evidence (β_{Corr} and β_{Incorr}). Note that in most variants of SDT models, a linear regression captures the relationship between confidence and evidence well, as long as confidence does not reach ceiling or floor values (*42*, *43*).

For display purposes (Figs. 2 and 4 to 6C), data were divided into six bins for each individual, incentive, and response (correct or incorrect) level. Scatterplots display the population-averaged data (and SEM) for each incentive and response (correct or incorrect) level.

Statistics

All statistical analyses were performed with MATLAB R2015a. All statistical analyses reported in the main text result from the linear mixed-effects models (estimated with the fitglme function). For each (nonreaction time) behavioral (for example, confidence and performance) and metacognitive (bias and sensitivity) measure Y, we computed the average of Y per incentive level per individual. For reaction times, whose distributions are typically skewed, we computed the median, rather than the mean, reaction time in each incentive condition. For the confidence formation model, we used the regression coefficient from the individual linear regressions linking confidence and evidence for correct and incorrect choices, estimated per individual and incentive level. We then used the absolute incentive value ($|V|$), the net incentive value (V), and the incentive valence (+/−, only for experiments 3 and 4) as predictor variables. All mixed models included random intercepts and random

slopes. As an example, in Wilkinson-Rogers notation, the linear mixed-effects models for experiment 1 can be written as follows: $Y \sim 1 + |V| + V + (1 + |V| + V\,|\text{Subject})$. Detailed results on all linear mixed-effects models used in the study can be found in Supplementary Results.

Deriving the cost of reporting biased confidence

To estimate the expected cost (in terms of winning probability) of a bias b, we can compute the difference between an expected win with and without this bias.

$$\text{Cost} = E(x+b) - E(x) \tag{11}$$

Using Eq. 7 derived in the "Optimal confidence rating in an MP elicitation mechanism" section, this gives

$$\text{Cost} = \left[-(x+b)^2 + 2(x+b)c + 1 - c\right] - \left[-x^2 + 2xc + 1 - c\right] \tag{12}$$

$$\text{Cost} = \left[-(x^2 + 2xb + b^2) + 2xc + 2bc + 1 - c\right] - \left[-x^2 + 2xc + 1 - c\right] \tag{13}$$

$$\text{Cost} = -(b^2 + 2b(x-c)) \tag{14}$$

There are several things worth noting.

First, this analytical approach allows us to estimate the pure effect of an additional bias. This is particularly important in our case, given that incentives also have a motivational effect on confidence accuracy.

Second, if $x = c$, that is, if the confidence rating before the bias was optimal, then the cost function of a bias b is $-b^2$, that is, a simple quadratic cost function.

Third, if confidence is already biased (for example, $x > c$ because individuals are overconfident), then the additional bias b combines with this existing bias and induces extra loss, because the loss function is not only additive but also quadratic (see also Fig. 6B).

Finally, in the specific case of the incentive bias demonstrated in the present report, the bias b is a function of incentives I

$$b = \beta_V \times I \tag{15}$$

where β_V is the unstandardized regression coefficient assessing the effect of net incentive value on confidence and I is the value of the incentive (in euros). We can then derive the additional expected monetary cost of this bias, in euros

$$\text{Monetary cost} = -\text{abs}(I) \times \left[(\beta_V \times I)^2 + 2(\beta_V \times I)(x-c)\right] \tag{16}$$

Note that this simple model is descriptive and was only developed to illustrate the consequences of the incentive bias in the context of the present setting. A full mechanistic model should include, for example, a boundary condition to make sure that biased confidence $(x + \beta_V \times I)$ remains a proper confidence judgment, that is, takes values between 50 and 100%.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/4/5/eaaq0668/DC1

REFERENCES AND NOTES

1. R. M. Cooke, *Experts in Uncertainty: Opinion and Subjective Probability in Science* (Oxford Univ. Press, 1992).
2. J. K. Adams, A confidence scale defined in terms of expected percentages. *Am. J. Psychol.* **70**, 432–436 (1957).
3. A. Pouget, J. Drugowitsch, A. Kepecs, Confidence and certainty: Distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19**, 366–374 (2016).
4. S. M. Fleming, R. J. Dolan, The neural basis of metacognitive ability. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 1338–1349 (2012).
5. N. Yeung, C. Summerfield, Metacognition in human decision-making: Confidence and error monitoring. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 1310–1321 (2012).
6. G. W. Brier, Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1–3 (1950).
7. S. M. Fleming, H. C. Lau, How to measure metacognition. *Front. Hum. Neurosci.* **8**, 443 (2014).
8. T. O. Nelson, A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychol. Bull.* **95**, 109–133 (1984).
9. T. Folke, C. Jacobsen, S. M. Fleming, B. De Martino, Explicit representation of confidence informs future value-based decisions. *Nat. Hum. Behav.* **1**, 0002 (2016).
10. F. Meyniel, D. Schlunegger, S. Dehaene, The sense of confidence during probabilistic learning: A normative account. *PLOS Comput. Biol.* **11**, e1004305 (2015).
11. M. Donoso, A. G. E. Collins, E. Koechlin, Foundations of human reasoning in the prefrontal cortex. *Science* **344**, 1481–1486 (2014).
12. F. Vinckier, R. Gaillard, S. Palminteri, L. Rigoux, A. Salvador, A. Fornito, R. Adapa, M. O. Krebs, M. Pessiglione, P. C. Fletcher, Confidence and psychosis: A neuro-computational account of contingency learning disruption by NMDA blockade. *Mol. Psychiatry* **21**, 946–955 (2016).
13. S. Lichtenstein, B. Fischhoff, L. D. Phillips, Calibration of probabilities: The state of the art to 1980, in *Judgment Under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic, A. Tversky, Eds. (Cambridge Univ. Press, 1982), pp. 306–334.
14. J. V. Baranski, W. M. Petrusic, The calibration and resolution of confidence in perceptual judgments. *Percept. Psychophys.* **55**, 412–428 (1994).
15. R. F. West, K. E. Stanovich, The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. *Psychon. Bull. Rev.* **4**, 387–392 (1997).
16. E. S. Berner, M. L. Graber, Overconfidence as a cause of diagnostic error in medicine. *Am. J. Med.* **121**, S2–S23 (2008).
17. C. Camerer, D. Lovallo, Overconfidence and excess entry: An experimental approach. *Am. Econ. Rev.* **89**, 306–318 (1999).
18. U. Malmendier, G. Tate, CEO overconfidence and corporate investment. *J. Finance* **60**, 2661–2700 (2005).
19. V. L. Smith, J. M. Walker, Monetary rewards and decision cost in experimental economics. *Econ. Inq.* **31**, 245–261 (1993).
20. S. E. Bonner, G. B. Sprinkle, The effects of monetary incentives on effort and task performance: Theories, evidence, and a framework for research. *Account. Organ. Soc.* **27**, 303–345 (2002).
21. N. T. Wilcox, Lottery choice: Incentives, complexity and decision time. *Econ. J.* **103**, 1397–1417 (1993).
22. M. Botvinick, T. Braver, Motivation and cognitive control: From behavior to neural mechanism. *Annu. Rev. Psychol.* **66**, 83–113 (2015).
23. Z. Kunda, The case for motivated reasoning. *Psychol. Bull.* **108**, 480–498 (1990).
24. R. Bénabou, J. Tirole, Mindful economics: The production, consumption, and value of beliefs. *J. Econ. Perspect.* **30**, 141–164 (2016).
25. N. Epley, T. Gilovich, The mechanics of motivated reasoning. *J. Econ. Perspect.* **30**, 133–140 (2016).

26. F. Giardini, G. Coricelli, M. Joffily, A. Sirigu, Overconfidence in predictions as an effect of desirability bias, in *Advances in Decision Making Under Risk and Uncertainty*, M. Abdellaoui, J. D. Hey, Eds. (Springer Berlin Heidelberg, 2008), pp. 163–180.
27. P. Koellinger, T. Treffers, Joy leads to overconfidence, and a simple countermeasure. *PLOS ONE* **10**, e0143263 (2015).
28. S. Massoni, Emotion as a boost to metacognition: How worry enhances the quality of confidence. *Conscious. Cogn.* **29**, 189–198 (2014).
29. F. U. Jönsson, H. Olsson, M. J. Olsson, Odor emotionality affects the confidence in odor naming. *Chem. Senses* **30**, 29–35 (2005).
30. M. Allen, D. Frank, D. S. Schwarzkopf, F. Fardo, J. S. Winston, T. U. Hauser, G. Rees, Unexpected arousal modulates the influence of sensory noise on confidence. *eLife* **5**, e18103 (2016).
31. B. De Martino, S. M. Fleming, N. Garrett, R. J. Dolan, Confidence in value-based choice. *Nat. Neurosci.* **16**, 105–110 (2013).
32. M. Lebreton, R. Abitbol, J. Daunizeau, M. Pessiglione, Automatic integration of confidence in the brain valuation signal. *Nat. Neurosci.* **18**, 1159–1167 (2015).
33. M. Guggenmos, G. Wilbertz, M. N. Hebart, P. Sterzer, Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife* **5**, e13388 (2016).
34. M. N. Hebart, Y. Schriever, T. H. Donner, J.-D. Haynes, The relationship between perceptual decision variables and confidence in the human brain. *Cereb. Cortex* **26**, 118–130 (2016).
35. D. J. Levy, P. W. Glimcher, The root of all value: A neural common currency for choice. *Curr. Opin. Neurobiol.* **22**, 1027–1038 (2012).
36. N. Schwarz, G. L. Clore, Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *J. Pers. Soc. Psychol.* **45**, 513–523 (1983).
37. G. M. Becker, M. H. Degroot, J. Marschak, Measuring utility by a single-response sequential method. *Behav. Sci.* **9**, 226–232 (1964).
38. W. M. Ducharme, M. L. Donnell, Intrasubject comparison of four response modes for "subjective probability" assessment. *Organ. Behav. Hum. Perform.* **10**, 108–117 (1973).
39. A. Schotter, I. Trevino, Belief elicitation in the laboratory. *Annu. Rev. Econ.* **6**, 103–128 (2014).
40. K. H. Schlag, J. Tremewan, J. J. van der Weele, A penny for your thoughts: A survey of methods for eliciting beliefs. *Exp. Econ.* **18**, 457–490 (2015).
41. G. Hollard, S. Massoni, J.-C. Vergnaud, In search of good probability assessors: An experimental comparison of elicitation rules for confidence judgments. *Theory Decis.* **80**, 363–387 (2016).
42. J. I. Sanders, B. Hangya, A. Kepecs, Signatures of a statistical computation in the human sense of confidence. *Neuron* **90**, 499–506 (2016).
43. S. M. Fleming, N. D. Daw, Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychol. Rev.* **124**, 91–114 (2017).
44. J. Drugowitsch, R. Moreno-Bote, A. Pouget, Relation between belief and performance in perceptual decision making. *PLOS ONE* **9**, e96511 (2014).
45. J. Drugowitsch, Becoming confident in the statistical nature of human confidence judgments. *Neuron* **90**, 425–427 (2016).
46. B. Maniscalco, H. Lau, A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cogn.* **21**, 422–430 (2012).
47. S. Massoni, T. Gajdos, J.-C. Vergnaud, Confidence measurement in the light of signal detection theory. *Front. Psychol.* **5**, 1455 (2014).
48. S. J. Galvin, J. V. Podd, V. Drga, J. Whitmore, Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychon. Bull. Rev.* **10**, 843–876 (2003).
49. A. Tversky, D. Kahneman, Loss aversion in riskless choice: A reference-dependent model. *Q. J. Econ.* **106**, 1039–1061 (1991).
50. R. F. Baumeister, Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *J. Pers. Soc. Psychol.* **46**, 610–620 (1984).
51. S. M. Fleming, R. J. Dolan, Effects of loss aversion on post-decision wagering: Implications for measures of awareness. *Conscious. Cogn.* **19**, 352–363 (2010).
52. J. Navajas, B. Bahrami, P. E. Latham, Post-decisional accounts of biases in confidence. *Curr. Opin. Behav. Sci.* **11**, 55–60 (2016).
53. S. T. Trautmann, G. van de Kuilen, Belief elicitation: A horse race among truth serums. *Econ. J.* **125**, 2116–2135 (2015).
54. P. P. Wakker, On the composition of risk preference and belief. *Psychol. Rev.* **111**, 236–241 (2004).
55. C. R. Fox, R. A. Poldrack, Prospect theory and the brain, in *Neuroeconomics: Decision Making and the Brain*, P. W. Glimcher, C. F. Camerer, E. Fehr, R. A. Poldrack, Eds. (Academic Press, 2009), pp. 145–173.
56. Z. Murad, M. Sefton, C. Starmer, How do risk attitudes affect measured confidence? *J. Risk Uncertain.* **52**, 21–46 (2016).
57. N. Kandasamy, S. N. Garfinkel, L. Page, B. Hardy, H. D. Critchley, M. Gurnell, J. M. Coates, Interoceptive ability predicts survival on a London trading floor. *Sci. Rep.* **6**, 32986 (2016).
58. C. Song, R. Kanai, S. M. Fleming, R. S. Weil, D. S. Schwarzkopf, G. Rees, Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Conscious. Cogn.* **20**, 1787–1792 (2011).
59. H. A. Harsay, M. X. Cohen, L. Reneman, K. R. Ridderinkhof, How the aging brain translates motivational incentive into action: The role of individual differences in striato-cortical white matter pathways. *Dev. Cogn. Neurosci.* **1**, 530–539 (2011).
60. H. Olsson, A. Winman, Underconfidence in sensory discrimination: The interaction between experimental setting and response strategies. *Percept. Psychophys.* **58**, 374–382 (1996).
61. P. Juslin, A. Winman, H. Olsson, Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychol. Rev.* **107**, 384–396 (2000).
62. D. D. P. Johnson, J. H. Fowler, The evolution of overconfidence. *Nature* **477**, 317–320 (2011).
63. J. A. R. Marshall, P. C. Trimmer, A. I. Houston, J. M. McNamara, On evolutionary explanations of cognitive biases. *Trends Ecol. Evol.* **28**, 469–473 (2013).
64. F. Meyniel, M. Sigman, Z. F. Mainen, Confidence as Bayesian probability: From neural origins to behavior. *Neuron* **88**, 78–92 (2015).
65. A. S. Goodie, The role of perceived control and overconfidence in pathological gambling. *J. Gambl. Stud.* **21**, 481–502 (2005).
66. R. Z. Goldstein, N. Alia-Klein, D. Tomasi, L. Zhang, L. A. Cottone, T. Maloney, F. Telang, E. C. Caparelli, L. Chang, T. Ernst, D. Samaras, N. K. Squires, N. D. Volkow, Is decreased prefrontal cortical sensitivity to monetary reward associated with impaired motivation and self-control in cocaine addiction? *Am. J. Psychiatry* **164**, 43–51 (2007).
67. M. Figee, M. Vink, F. Geus, N. Vulink, D. J. Veltman, H. Westenberg, D. Denys, Dysfunctional reward circuitry in obsessive-compulsive disorder. *Biol. Psychiatry* **69**, 867–874 (2011).
68. A. E. Whitton, M. T. Treadway, D. A. Pizzagalli, Reward processing dysfunction in major depression, bipolar disorder and schizophrenia. *Curr. Opin. Psychiatry* **28**, 7–12 (2015).
69. S. Yu, T. J. Pleskac, M. D. Zeigenfuse, Dynamics of postdecisional processing of confidence. *J. Exp. Psychol. Gen.* **144**, 489–510 (2015).
70. R. van den Berg, K. Anandalingam, A. Zylberberg, R. Kiani, M. N. Shadlen, D. M. Wolpert, A common mechanism underlies changes of mind about decisions and confidence. *eLife* **5**, e12192 (2016).

Acknowledgments: We thank S. Palminteri, I. Soraperra, F. van Winden, J. B. Engelmann, and J. van der Weele for helpful discussions and comments on the manuscript and T. A. Davison for checking the English. **Funding:** M.L., J.L., and R.J.v.H. were supported by individual Amsterdam Brain and Cognition Talent Grants (Universiteit van Amsterdam). M.L. was additionally supported by an NWO Veni Fellowship (grant 451-15-015) and the Bettencourt Schueller Fondation. A.E.G. received funding through an NWO Vidi scheme (grant 91713354) and an Aspasia grant from The Netherlands Organisation for Health Research and Development (NWO-ZonMw). **Author contributions:** M.L., R.J.v.H., and J.L. designed the study. S.L., M.J.S., and J.S.N. collected data. M.L. analyzed data. A.E.G. and D.D. provided supervision. M.L., R.J.v.H., and J.L. wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. All codes and data needed to evaluate or reproduce the figures and analysis described in the paper are available online at https://dx.doi.org/10.6084/m9.figshare.6126776.

Submitted 27 September 2017
Accepted 18 April 2018
Published 30 May 2018
10.1126/sciadv.aaq0668

Citation: M. Lebreton, S. Langdon, M. J. Slieker, J. S. Nooitgedacht, A. E. Goudriaan, D. Denys, R. J. van Holst, J. Luigjes, Two sides of the same coin: Monetary incentives concurrently improve and bias confidence judgments. *Sci. Adv.* **4**, eaaq0668 (2018).

ScienceAdvances

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

Maël Lebreton, Shari Langdon, Matthijs J. Slieker, Jip S. Nooitgedacht, Anna E. Goudriaan, Damiaan Denys, Ruth J. van Holst and Judy Luigjes

Sci Adv **4** (5), eaaq0668.
DOI: 10.1126/sciadv.aaq0668

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.