Title:

Deciphering the effects of incentive motivation on probabilistic judgments.

Authors

Nahuel Salem-Garcia[1,2], Sébastien Massoni[3]* and Maël Lebreton[2,4,5]*

Affiliations

[1]LNC2, Département d'études cognitives, Ecole normale supérieure, Université PSL, INSERM, 75005 Paris, France

[2]Paris School of Economics, 48 Bd Jourdan, 75014 Paris, France

[3]Université de Lorraine, Université de Strasbourg, CNRS, BETA, Nancy, France

[4]CNRS UMR8545, PjSE, 48 Bd Jourdan, 75014 Paris, France

[5]Swiss Center for Affective Science, Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland

* These authors contributed equally

Corresponding authorssalemnahuel@gmail.com; mael.lebreton@psemail.eu

Keywords:

Beliefs, confidence, incentive motivation, biases, elicitation mechanisms

Authors contributions

NSG, SM and ML designed the study. SM acquired funding. NSG performed the experiments and analyzed the data. NSG, SM and ML interpreted the analyses. NSG and ML drafted the manuscript. All authors critically reviewed the manuscript.

Acknowledgement

The study was supported by research funds allocated to SM (programme FUTURE LEADER of Lorraine Université d'Excellence within the French National Research Agency *Investissements d'avenir* ANR-15-IDEX-04-LUE). ML acknowledges the support of the European Research Council (Starting Grant INFORL-948671). NSG was also partly supported by the Swiss National Science Foundation (Postdoc.Mobility: P500PS_214314) and the Agence Nationale de la Recherche (MONODEC: ANR-23-CE37-0028-02). PSE is supported by the French National Research Agency (Investissements d'Avenir, ANR-10–LABX-93-0 and ANR-17–EURE-0001). The

funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Abstract (150/150 words):

The ubiquity of over-optimism, overconfidence, wishful thinking and desirability biases suggest that a universal motivational mechanism distorts all decisions and judgments uniformly. Here, we investigate this intuition, by assessing the relative effects of incentives on two types of probabilistic judgments: beliefs about externally generated hypotheses, versus confidence in one's own decisions. Across four perceptual decision-making experiments, we manipulate the agency of decisions over ambiguous states-of-the-world and the monetary incentives for accurate probabilistic judgments about decision accuracy. Results show that incentives consistently bias participants' reports, but, contrary to the uniform motivational bias intuition, substantially more for confidence than belief judgments. Modelling probabilistic judgments as an incentive-dependent weighting of evidence that is congruent with a (covert) decision rationalizes this apparent discrepancy. We conclude that gain versus loss prospects modulate how we integrate evidence confirming our endorsed hypothesis, shedding light on the formation of beliefs and confidence, and their interaction with motivational processes.

Introduction

Beliefs, i.e. the subjective assessments of the probabilities of various types of hypotheticals and states-of-the-world, are key to the computations of expectations and therefore central to most decision-making theories (De Finetti, 1937; Jeffrey, 1990; Manski, 2004; Ramsey, 1931; Savage, 1954). Although the field of economics generally commit to this very broad definition of beliefs, cognitive psychology makes a specific distinction between first-order beliefs, i.e. probability judgments over (external) states, and second-order beliefs, i.e. probability judgments over one's own actions or statements being correct (Fleming & Daw, 2017; Pouget et al., 2016). This last category of belief is actually part of metacognitive processes –i.e. processes that assess other cognitive processes (Fleming & Dolan, 2012; Yeung & Summerfield, 2012)– and is formally referred to as confidence judgments (Fleming & Daw, 2017; Harvey, 1997; Pouget et al., 2016).

Because first-order beliefs and confidence judgments take the form of subjective probabilistic assessments, they are usually not directly observable (unlike preference). However, they can be properly elicited and compared to objective frequencies to scientifically and quantitatively investigate the accuracy of those judgments (Ducharme & Donnell, 1973; Hollard et al., 2016; Karni, 2009; Manski, 2004; Schlag et al., 2015; Schotter & Trevino, 2014). Using these theoretical (formalisation as probabilistic judgments) and experimental (elicitation methods) tools, decades of empirical research in economics and psychology have revealed that, much like all other aspects of decision making, our beliefs seem to be susceptible to numerous biases and framing effects (Lichtenstein et al., 1982; Shekhar & Rahnev, 2021; Wallsten & Budescu, 1983).

Among the most famous and prominent of such biases, one class is of particular interest for the present study: motivational and affective biases. In short, probabilistic judgments about future states (beliefs) or about the correctness of decision, statement or actions (confidence) are affected by the decision-maker's preferences and goals. Practically, we end up overestimating the probability of desirable events, a phenomenon referred to as (over)optimism or wishful thinking (Babad & Katz, 1991; Sharot, 2011; Van den Steen, 2004). We also have a tendency to overestimate the probability that our decisions and statements are correct, a bias referred to as overconfidence (Fischhoff et al., 1977; Moore & Healy, 2008). Both biases are especially pronounced in positive contexts and moods,

and when facing positive expected-value prospects (Giardini et al., 2008; Koellinger & Treffers, 2015; Lebreton et al., 2018; Massoni, 2014).

Overall, these empirical regularities seem to suggest that a general *desirability* or *motivational* bias uniformly affects beliefs and confidence judgments (Epley & Gilovich, 2016; Krizan & Windschitl, 2007; Kunda, 1990). This idea is supported by influential theories of belief-based utility from economics. Individuals might choose to think that future prospects are brighter than they are, or that their own decisions or statements are better or more correct than they are. Despite being incorrect, holding those beliefs has some consumption, signalling or motivational value. Thereby; they bear psychological or strategic benefits, such as alleviating anxiety, improving ego-perception and positive self-signaling (Bénabou & Tirole, 2002; Eil & Rao, 2011; Loewenstein & Molnar, 2018).

Here, building on a recently developed experimental framework (Lebreton et al., 2018), we systematically manipulate the agency of categorical decisions about ambiguous percepts –creating identical situations framed either as confidence or belief judgments (Pereira et al., 2020)–, the monetary incentives for accurate probabilistic judgments about the correctness of those decisions, and the incentivization scheme (i.e. the expected gain conditional on holding certain beliefs). These manipulations allow us to test the predictions of different hypotheses regarding the effects of incentives on probabilistic judgments. Across four experiments (n = 800 participants, 192,000 total observations; N = 780 participants, 186,716 observations after exclusions, see **Methods**), we identify and replicate a clear and specific effect of the net incentive value on confidence which almost vanishes when considering beliefs. Thanks to original experimental manipulations, we rule out the possibility that this discrepancy is due to the incentivization procedure, to the motor aspect of the decision, or to the fact that the decision was produced by the agent. Taking these results as an indication that belief accuracy incentives affect the mechanism that link together the integration of stimuli, the (latent) decision and the probabilistic judgments, we propose a simplified model based on the principle of an incentive-sensitive balance between choice-congruent and -incongruent evidence. Deciphering the mechanisms governing the formation of beliefs and confidence judgments and their interaction with motivational and affective processes appears critical to assess the consequences of various framing of probabilistic judgment elicitation mechanisms and appraise the limits of popular behavioral theories (motivated beliefs, desirability bias).

Results

Experiment 1: Incentive motivation differently affects beliefs and confidence judgments

Two hundred participants took part in our first experiment (Exp 1a; N= 196 after exclusions, see ***Methods***). All participants performed a version of our task that elicits, over multiple trials, incentivized probabilistic judgments on categorical decisions about ambiguous, perceptual stimuli. Thereby, after briefly seeing a random-dot kinematogram (RDK), participants judged the probability of a decision about the main direction of moving dots being correct (see **Methods** and **Figure 1A**). We implemented two key treatments: our first, within-participant treatment varied the incentive offered at each trial for the accurate probabilistic judgment; thereby, depending on the trial, participants could earn £1 (gain condition), nothing (neutral condition) or avoid losing £1 (loss condition) for accurate probabilistic judgment −see (Hoven, Brunner, et al., 2022; Lebreton et al., 2018) for a similar design and **Figure 1A** for details. Our second, between-participant treatment varied the agency of the decisions whose correctness had then to be estimated: a hundred participants made the decisions and a hundred participants observed decisions made by the computer. We refer to the first condition, where the probability judgment corresponds to a confidence judgment, as the free-choice condition, and to the second condition, where the probability judgment corresponds to a belief judgment because participants did not make the decision, as the observed-choice condition. Note that the exact same distribution of difficulty (i.e. perceptual evidence) was used across treatments, and that the correctness of the computer's decisions conditional on the difficulty matched the one observed in actual participants in a pilot study (see **Methods**). In all treatments, the probability of being correct was elicited as a 0-100% rating and an incentivization mechanism treated this judgment as a bet against a random lottery with 0-100% support –a so-called Probability Matching mechanism (Ducharme & Donnell, 1973; Hollard et al., 2016; Holt & Smith, 2009; Karni, 2009; Schlag et al., 2015; Schotter & Trevino, 2014). In order to identify the effects of affective and motivational biases on beliefs and confidence judgments, we ran hierarchical regression models on several key behavioural variables, that jointly modelled the effects of the incentive net value V and of the incentive absolute value |V|. These two orthogonal variables conveniently capture the affective value of incentives (that takes negative values in the loss framing and positive in the gain framing) and their motivational aspects (that takes positive value when money is at stake).

We first inspected self-generated and observed decision accuracy, i.e. how well choices (self-generated or observed) matched the main direction of the RDK (**Figure 1A**, and see **Methods** for details about how this was computed). As expected from our design, decision accuracy did not differ across our agency and incentive treatments (**Figure 1B, Supp. Table 2**). This initial check confirms that we can safely analyse and contrast the effects of our treatments on the reported probability of the choice being correct (thereafter referred to as p(correct)) without any risk of a decision (i.e. type-1) accuracy confound. Note, however that all analyses reported below replicated when performed on calibration, i.e. on the average difference between the reported p(correct) and the actual decision-accuracy (see **Supp. Tables 1-2**). In the free-choice condition, similarly to Lebreton et al. (2018), we found a clear effect of the net incentive value, and not of the absolute incentive value on p(correct) (β_V = 3.45 ±0.53, $t_{94.95}$ = 6.50, P < 0.001, $\beta_{|V|}$ = 0.17 ±0.52, $t_{94.85}$ = 0.33, P = 0.740; **Figure 1B** and **Supp. Tables 1-2**). This effect replicated, but with a much lower effect-size in the observed-choice condition (β_V = 1.05 ±0.25, $t_{118.06}$ = 4.27, P < 0.001; $\beta_{|V|}$ = 0.32 ±0.40, $t_{511.27}$ = 0.80, P = 0.425), such that the effects of V were significantly higher in the free-choice ($\beta_{V:FREE}$ = 2.39 ±0.58, $t_{193.77}$ = 4.16, P < 0.001).

To verify that the effects identified in our linear regression are not driven by one framing, we also evaluated single contrasts between the incentivized conditions (gain and loss framing) and the neutral condition. This analysis confirmed that the effect of V is caused by a joint positive effect of gains and negative effect of losses on the p(correct) (**Supp. Tables 3**).

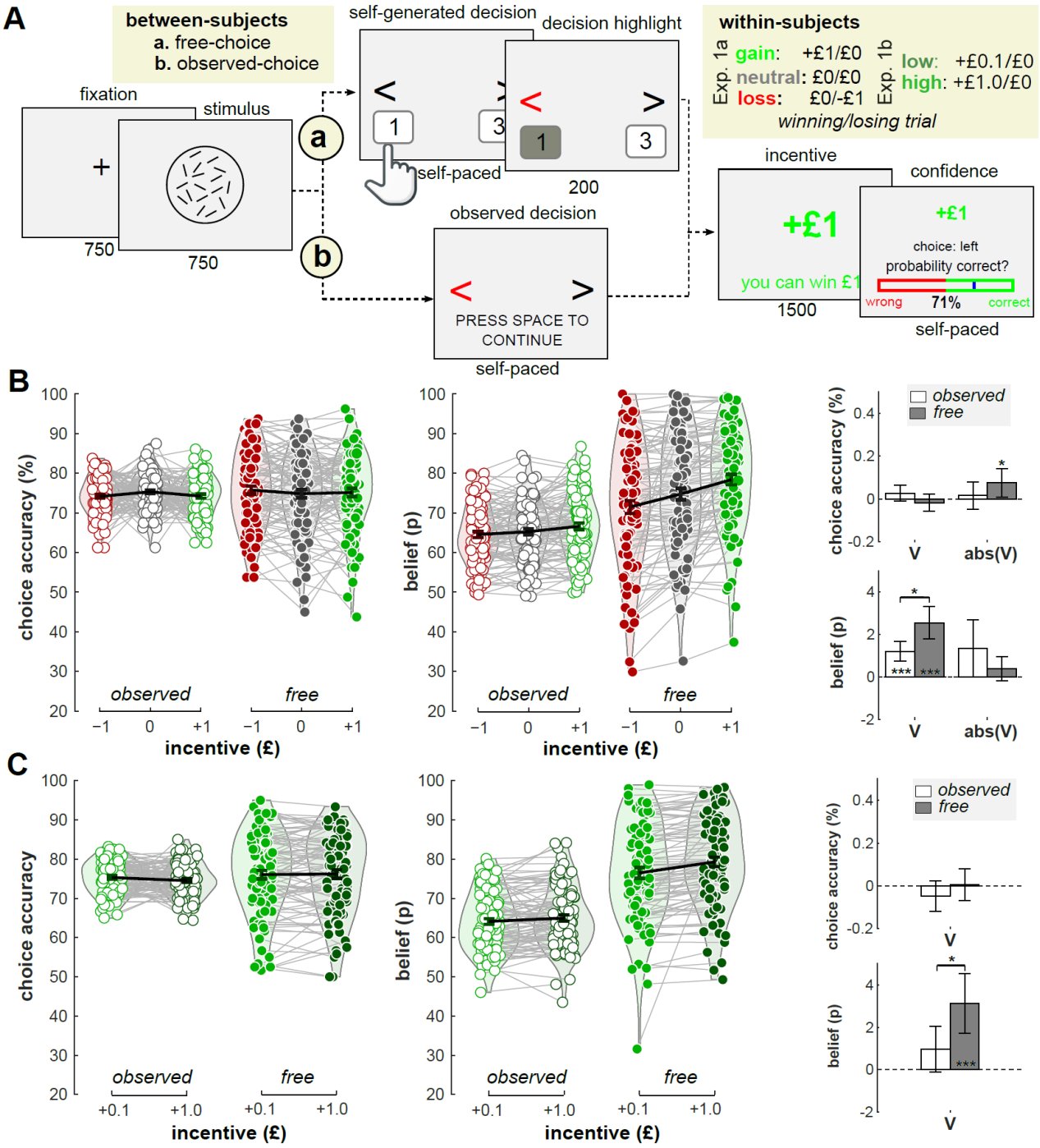

Figure 1 | **Experiment 1**. A. Example trial for free-choice (a) and observed-choice (b) conditions. Successive screens displayed in one trial are shown from left to right with durations in milliseconds B. Results of Experiment 1a. C. Results of Experiment 1b. Left panels picture the sample and individual-level distributions of the behavioural variable of interest (accuracy, p(correct), split by incentive (Exp 1a: gain: green; neutral: grey; loss: red; Exp 1b: low: dark green; high: light green) and agency conditions. Within the violin plots, the white dot represents the sample mean, the error bar indicates the mean ± 95%CI and light-colored dots represent all individual datapoints. The right panels picture the coefficient estimates from mixed-effect regression, corresponding to the effect of the net incentive value (V) and the absolute incentive value (|V|) on the behavioural variables (logistic regression for accuracy, linear for belief), in the two agency conditions (free-choice: white dot, observed-choice: grey dot). The error bar indicates the mean of the coefficient estimate ± 95%CI.

*** p < 0.001, ** p < 0.01, * p < 0.05, # p < 0.10 (two-tailed tests).

We replicated and generalized those results in a second experiment (Exp. 1b), where we manipulated the magnitude of incentives (low-value: 10c; and high-value: 1€) in lieu of the framing (gain, neutral and loss). We hypothesized that, if the motivational bias on probabilistic judgment is due to the net incentive value, participants should report higher p(correct) in the high-magnitude condition. As a second, between-subject treatment, we again varied the agency of the decisions whose correctness had then to be estimated (free-choice versus observed-choice). Again, two samples of 100 participants took part in our between-subject design (N = 192 after exclusions, see **Methods**). Our analyses replicated the effect of the net incentive value on the reported p(correct) in free-choice conditions (β_V = 3.13 ± 0.72, $t_{91.01}$ = 4,34, P < 0.001), that was smaller in the observed choice condition (β_V = 0.97 ± 0.55, $t_{98.90}$ =1.76, P = 0.082; $\beta_{V:FREE}$ = 2.16 ± 0.90, $t_{190.02}$ = 2.41, P = 0.017; see **Figure 1C** and **Supp. Tables 4-5)**. These results suggest that the magnitude of incentives has a similar effect as their valence, hence supporting the hypothesis that the biasing effect of incentives are due to the net value.

Overall, the analyses of Exp.1a and Exp.1b coherently point towards two core results. First, the incentive effects on p(correct) manifest as a specific effect of their affective value (V), with no concurrent detectable effect of their motivational value (|V|). To interpret this first result, recall that the incentivization mechanism makes truthful reporting of one's belief about the probability of being correct the optimal strategy. However, the higher this probability, the greater the chance of a winning trial. Consequently, both in gain and loss conditions, higher beliefs translate into higher expected values (**Figure 2A**). Therefore, our first core result contrasts with a direct prediction of a general desirability bias: if participants prefer to hold beliefs about advantageous states of the world, this should manifest as a positive effect of the incentive absolute value |V| on the reported the probability of being correct. This result also contrasts with a rational effect of incentives. Our second core result lies in the finding that the identified biasing effects of incentives are much more substantial when probabilistic judgments are framed as a confidence judgment than as a belief judgment, despite the remarkable similarity between those two framings. In the remaining of this manuscript, we therefore attempted to consolidate and explain this puzzling pattern.

Experiment 2 – controlling for correlation between choice accuracy and probabilistic judgments.

The most evident difference between our conditions is the inherent higher correlation between the quality of the perceptual decision and the reported subjective probability of the choice being correct in the free-choice versus the observed-choice condition. Indeed, in the free-choice condition, participants systematically make perceptual choices that they believe are correct. Although our observed-choice condition was designed to match the average accuracy of those choices, it breaks the trial-by-trial correlation between choices and judgments: there can be trials where participants are particularly attentive, and where the decision appears unambiguously incorrect. This had two major consequences: first, the average level of p(correct) in Exp. 1a-b was generally higher in the free-choice than in the observed-choice condition. Second, and consequently, the expected values of the holding accurate beliefs (or of distorting those beliefs) were different in those conditions, given that this is a convex function of the true p(correct) (**Figure 2A**).

To alleviate those concerns, we ran a second experiment where we modified the reporting scale from 0-100% as two subscales that allowed to report the probabilistic judgment as 50-100% bets on either a wrong or correct choice (**Figure 2B**). This allowed participants, notably in the observed choice condition, to report high confidence for observed choice that they believed were incorrect. We also adjusted the incentivization mechanism, so as to treat this new rating as a bet against a random lottery with support 50-100%, which addressed the difference in expected values for correct vs incorrect choices incurred by the previous framing (**Figure 2B**). Again, we implemented the free-choice and observed-choice condition as a between-subject design, for which we invited 200 participants (N = 196 after exclusions, **see Methods**).

Keeping in line with our analysis strategy, we first verified that there was no effect of our treatment on the free or observed choice accuracy (**Figure 2D**). Importantly, though, we found again a significant effect of incentive net value on the reported probability in both conditions, that was significantly stronger in the free-choice than in the observed choice conditions (Free: $\beta_V = 1.74 \pm 0.25$, $t_{94.97} = 7.00$, $P < 0.001$; Observed: $\beta_V = 1.01 \pm 0.15$, $t_{99.12} = 6.79$, $P < 0.001$; $\beta_{V:FREE} = 0.73 \pm 0.29$, $t_{193.98} = 2.55$, $P = 0.012$; **Supp. Table 6-8**). This confirmed that, across different versions of the reporting scale and of the incentivization scheme, incentives have a stronger impact on probability judgments that qualifies a freely made than an observed choice.

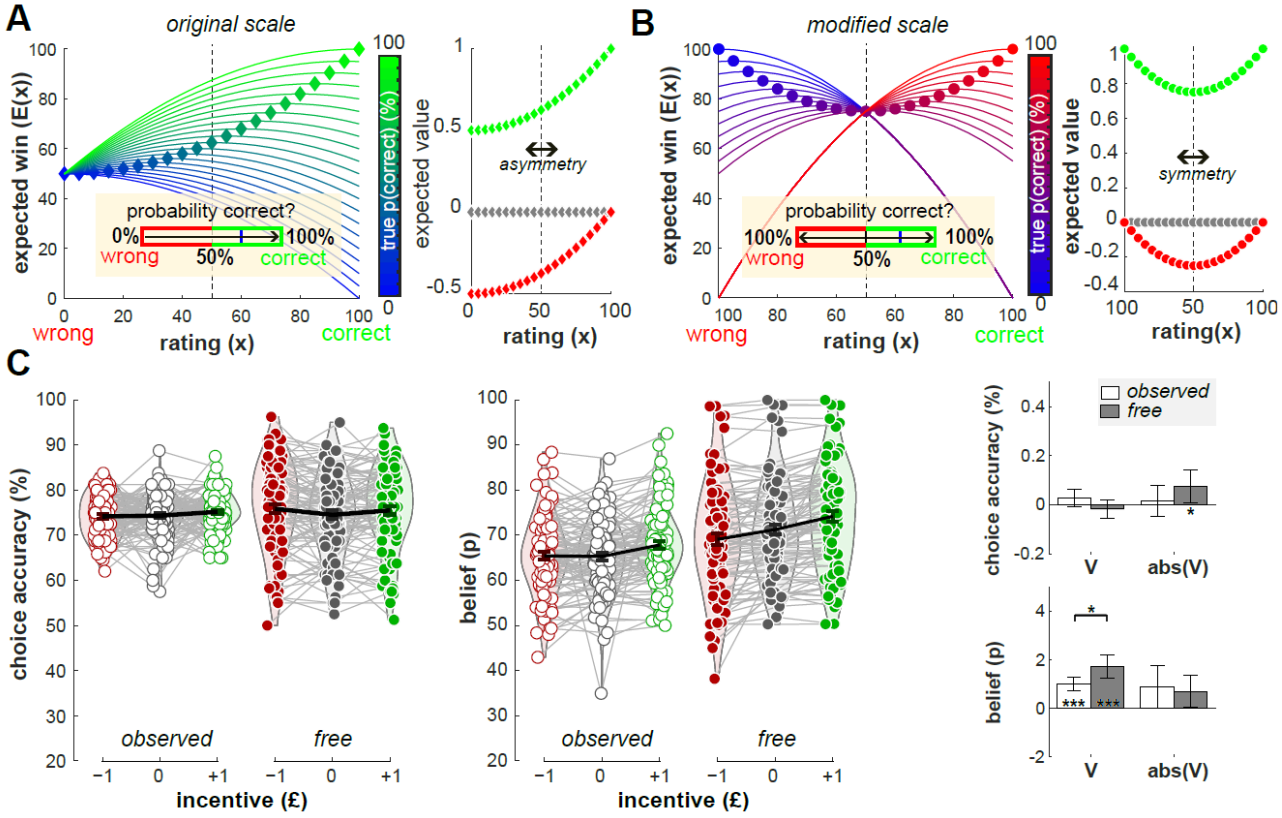

Figure 2 | Modifications for experiment 2. A-B. Incentivization mechanism for Experiment 1 (A) and 2 (B). The left panel pictures the expected probability of winning E(x) induced by the MP mechanism for several levels of subjective belief (true p(correct); depicted with the blue-green gradient), as a function of the chosen rating x. The diamonds indicate the highest point of each curve. One can visually check that for each level of belief, the rating that maximizes the expected probability of winning is the truthful reporting of this belief. The right panel showcase the expected payoff corresponding to each subjective belief, if truthfully reported, for gain, neutral and loss conditions (respectively green, grey and red). C. Results of Experiment 2. Left panels picture the sample and individual-level distributions of the behavioural variable of interest (accuracy, p(correct)), split by incentive (gain: green; neutral: grey; loss: red) and agency conditions. Within the violin plots, the white dot represents the sample mean, the error bar indicates the mean ± 95%CI and light-colored dots represent all individual datapoints. The right panels picture the coefficient estimates from mixed-effect regression, corresponding to the effect of the net incentive value (V) and the absolute incentive value (|V|) on the behavioural variables (logistic regression for accuracy, linear for belief), in the two agency conditions (free-choice: white dot, observed-choice: grey dot). The error bar indicates the mean of the coefficient estimate ± 95%CI.
*** P < 0.001, ** P < 0.01, * P < 0.05, # P < 0.10 (two-tailed tests).

Experiment 3 – testing the elements of the choice that generate incentive bias

So far, our results clearly indicate that incentives have a stronger influence on probabilistic judgements about self-generated choices – i.e. confidence judgments. To better understand the potential mechanism behind this effect, we next attempted to identify what cognitive process, present in the free-choice condition and absent in the observed-choice condition, is critical to the emergence of the full-magnitude incentive bias. We propose to dissect choice into two components: the (abstract) decision and commitment to an option, versus the motor execution of the decision – which relates to the feeling of agency (Charles et al., 2020; Siedlecka et al., 2021; Wen et al., 2023). We therefore designed a new task, where we separately added to the basic observed-choice condition, the motor execution component on the one hand, and the (abstract) decision and commitment to an option on the other hand. This respectively resulted in a forced-choice condition, in which participants had to actively select an option already picked and identified by a mechanism identical to the observed condition, and a replayed condition, where participants performed binary choices in a first session that they then had to evaluate as in an observed condition in a second session (**Figure 3A**). Our reasoning was that we could then evaluate whether these treatments, built on top of the observed-choice condition, could restore an effect of incentive comparable to the free-choice condition. We recruited 100 participants per condition (N = 196 after exclusions, see **Methods**). In both the forced-choice condition and the replayed-choice condition, we found a small effect of net incentive values on reported p(correct) (forced-choice: $\beta_V = 1.55 \pm 0.34$, $t_{96.89} = 4.57$, $P < 0.001$; replayed-choice: $\beta_V = 1.30 \pm 0.23$, $t_{97.40} = 5.60$, $P < 0.001$; **Figure 3B** and **Supp. Tables 9-11**), comparable to the one previously assessed in the observed-choice condition (comparison of marginal incentive trends: $\Delta_{FORCED - OBSERVED} = 0.53 \pm 0.63$, $t_{823.80} = 0.84$, $P = 1$; $\Delta_{REPLAYED - OBSERVED} = 0.23 \pm 0.57$, $t_{55053.97} = 0.39$, $P = 1$) , and significantly inferior to the one assessed in the free-choice condition ($\Delta_{FREE - FORCED} = 1.66 \pm 0.58$, $t_{53922.72} = 2.87$, $P = 0.016$; $\Delta_{FREE - REPLAYED} = 1.97 \pm 0.64$, $t_{799.20} = 3.09$, $P = 0.01$; **Figure 3B** and **Supp. Tables 12-13**). This result suggests that neither the (abstract) decision and commitment to an option, nor the motor aspect of the choice are likely to be the main channel through which incentives bias the production of probabilistic judgments. Rather, our collection of results established thus far points toward the idea that the bias emerges at the heart of the process that connects the integration of evidence with the decision that has to be evaluated. In our last section, we propose a computational account

that leverages this principle to propose a tentative coherent explanation for this collection of results.

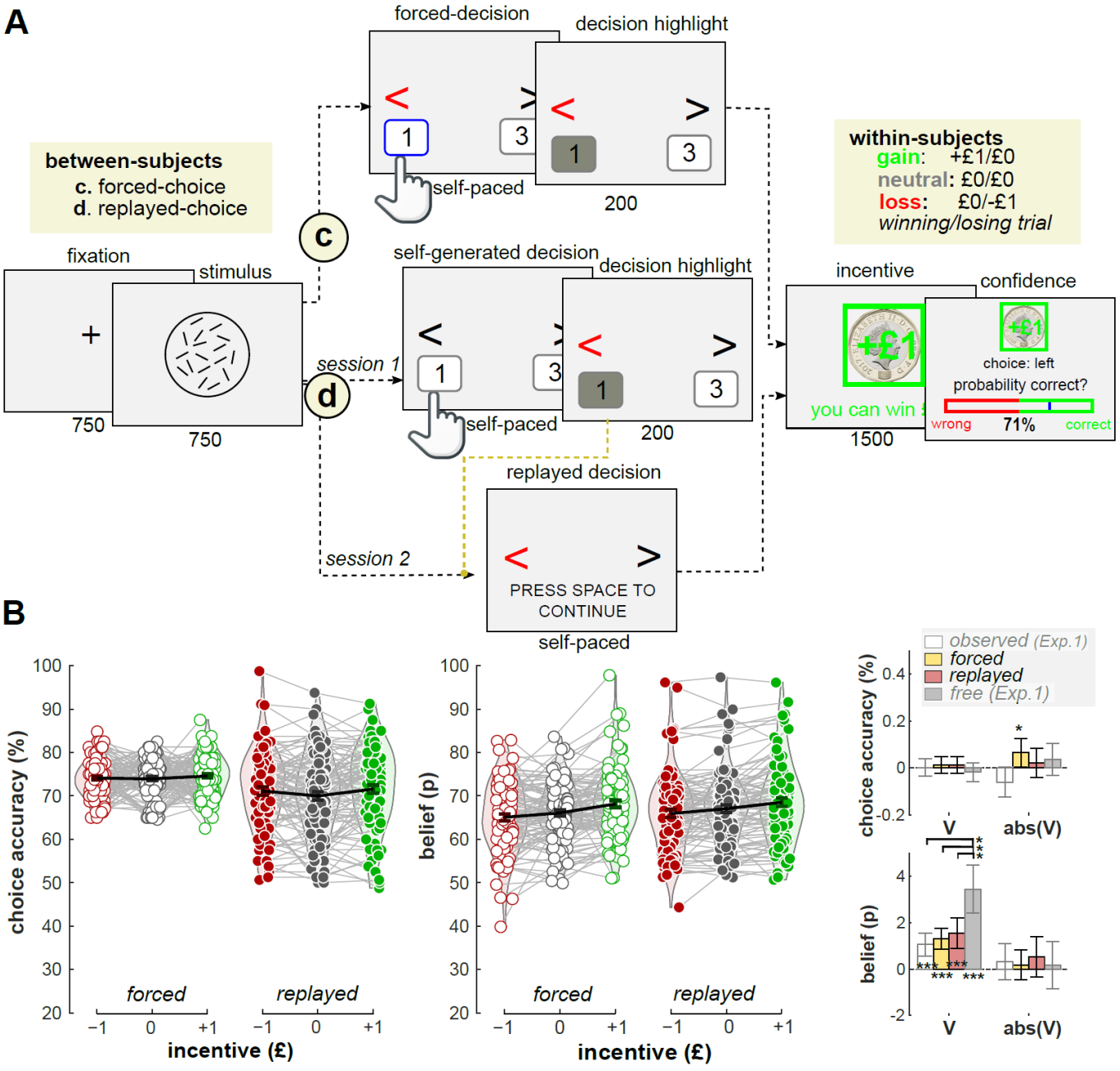

Figure 3 | Experiment 3. A. Example trial for forced-choice conditions. Successive screens displayed in one trial are shown from left to right with durations in milliseconds. Results. Left panels picture the sample and individual-level distributions of the behavioural variable of interest (accuracy, p(correct)), split by incentive (gain: green, neutral: grey; loss: red). Within the violin plots, the white dot represents the sample mean, the error bar indicates the mean ± 95%CI and light-colored dots represent all individual datapoints. The right panels picture the coefficient estimates from mixed-effect regression, corresponding to the effect of the net incentive value (V) and the absolute incentive value (|V|) on the behavioural variables (logistic regression for accuracy, linear for belief). The error bar indicates the mean of the coefficient estimate ± 95%CI.
*** P < 0.001, ** P < 0.01, * P < 0.05, # P < 0.10 (two-tailed tests).

Comparing models of incentive effects on probabilistic judgments

To provide a comprehensive account of our results, we built a family of three models, based on the increasingly consensual assumption that perception, decision, metacognitive judgments and all related latent variables and computations that are based on noisy percepts, which means that agents sometimes make wrong inference based on misperceived evidence (Fleming & Daw, 2017; Maniscalco & Lau, 2012; Sanders et al., 2016). Thereby, our computational models are grounded in 2D Signal Detection Theory and assume that participants sample noisy evidence for both options (left and right movement), make a covert decision, realize the corresponding action when allowed, and form a Bayesian posterior belief of the action (free or observed) being correct (see ***Methods*** for details). Then, in line we recent developments in the metacognition literature, our models adopt the principle that probabilistic judgments asymmetrically weight choice-congruent and choice-incongruent evidence, the latter being down-weighted or even discarded (Miyoshi & Lau, 2020; Peters et al., 2017; Salem-Garcia et al., 2023; Ting et al., 2023; Zylberberg et al., 2012). Critically, for the observed choices condition, this choice-congruency principle can be implemented in many different ways, three of which we considered in our model family (**Figure 4A**): our Action-Congruent model, is essentially a parametrically modulated extension of the heuristic Response-Congruent bias mentioned in the literature, and ignores the source of the response. As such, it behaves similarly in Free-choice and Observed conditions. On the other hand, our Intention-Congruent model exaggerates evidence towards the *covert* decision, i.e. the agent's internally inferred direction (Balsdon et al., 2021), instead of the response. When the observed decision/action is not-aligned with this covert decision, this model assumes that the agent reported the complementary probability. Finally, our Confirmation-Congruent model provides an intermediary account, as it exaggerates evidence towards the executed action, but only if it confirms the covert decision. In all our models, monetary incentives modulate the balance between congruent and incongruent evidence, as framed by the considered model. This naturally arises from the realization that gain-framing increases the relative focus on the proposition being correct ("you can win 1 euro if correct"), while the loss-framing shift this relative focus on the proposition being incorrect ("you can lose 1 euro if incorrect") – see (Sakamoto & Miyoshi, 2024) for a similar rationale.

Because those models essentially differ in how they link or decorrelate belief formation and (performed or observed) choices, they should make qualitatively different predictions on the relation between belief and choice accuracy, under the different incentive conditions (Fleming & Daw, 2017). We therefore evaluated the relative merit of our models by assessing how patterns of simulated beliefs for correct and wrong choices, in the different incentive conditions, qualitatively matched those observed in our participants behavior (Palminteri et al., 2017; **Figure 4B-E**, **Methods**, and **Supplementary Tables 14-15** for details of the simulation). In our data, the incentive effect is positive and present in both agency conditions (Free: $\beta_V = 3.00 \pm 0.33$, $t_{191} = 9.11$, $P < .001$, Observed: 1.30 ± 0.16, $t_{295} = 8.18$, $P < .001$), and higher in Free vs Observed ($\Delta_{FREE - OBSERVED} = 1.70 \pm 0.37$,$t_{280} = 4.66$, $P < .001$). This holds both in correct (Free: $\beta_{V|CORRECT} = 2.97 \pm 0.33$, $t_{191} = 9.08$, $P < .001$; Observed: $\beta_{V|CORRECT} = 1.30 \pm 0.17$, $t_{295} = 7.61$, $p < 0.001$; difference: $\Delta_{FREE - OBSERVED} = 1.67 \pm 0.37$, $t_{295} = 4.53$, $P < .001$) and incorrect choices (Free: $\beta_{V|INCORRECT} = 3.33 \pm 0.40$, $t_{191} = 8.33$, $P < .001$; Observed: $\beta_{V|INCORRECT} = 1.09 \pm 0.29$, $t_{295} = 3.79$, $P < .001$, difference: $\Delta_{FREE - OBSERVED} = 2.23$, $t_{377} = 4.53$, $P < .001$). Simulations first confirm that the Action-Congruent model can reproduce the presence of an incentive effect in both Free and Observed choices (Free: $\beta_V = 3.15 \pm 0.06$, $t_{191} = 54.08$, $p < .001$, Observed: 3.06 ± 0.07, $t_{191} = 43.12$, $P < .001$), but because it is agnostic to the source of the response, it predicts a similar magnitude of the effect between these ($\Delta_{FREE - OBSERVED} = 0.09 \pm 0.09$, $t_{368} = 0.96$, $P = .336$). The Intention-Congruent model exaggerates evidence towards the covert decision instead of the action, and therefore predicts an inversion of the incentive effect in trials where the covert decision mismatches action. This model therefore correctly accounts for the overall reduced incentive effect on beliefs ($\beta_V = 0.57 \pm 0.05$, $t_{191} = 11.66$, $P < .001$; $\Delta_{FREE - OBSERVED} = 2.58$, $t_{370} = 34.04$, $P < .001$). However, simulations show that it also incorrectly generates an inversion of the effect in incorrect Observed choices (increased beliefs for losses, decreased for gains; $\beta_{V|INCORRECT} = -1.02 \pm 0.09$, $t_{191} = -11.96$, $P < .001$). Finally, the Confirmation-Congruent model which exaggerates evidence towards the action, but only if it confirms the covert decision effectively negates the incentive bias in trials where the observed action mismatches the covert decision. In our simulations, this model does replicate the qualitative patterns observed in the data: the incentive effect is reduced without inversion in observed decisions, both in correct ($\beta_{V|CORRECT} = 1.98 \pm 0.05$, $t_{191} = 39.61$, $P < 0.001$; $\Delta_{FREE - OBSERVED} = 0.90 \pm 0.08$, $t_{380} = 11.88$, $P < .001$) and incorrect choices ($\beta_{V|INCORRECT} = 1.20 \pm 0.09$, $t_{191} = 13.85$, $P < .001$; $\Delta_{FREE - OBSERVED} = 2.86$, $t_{373} =$

21.61, P < .001). We conclude that this Confirmation-Congruent model can parsimoniously account for the qualitative patterns observed in our data with few assumptions.

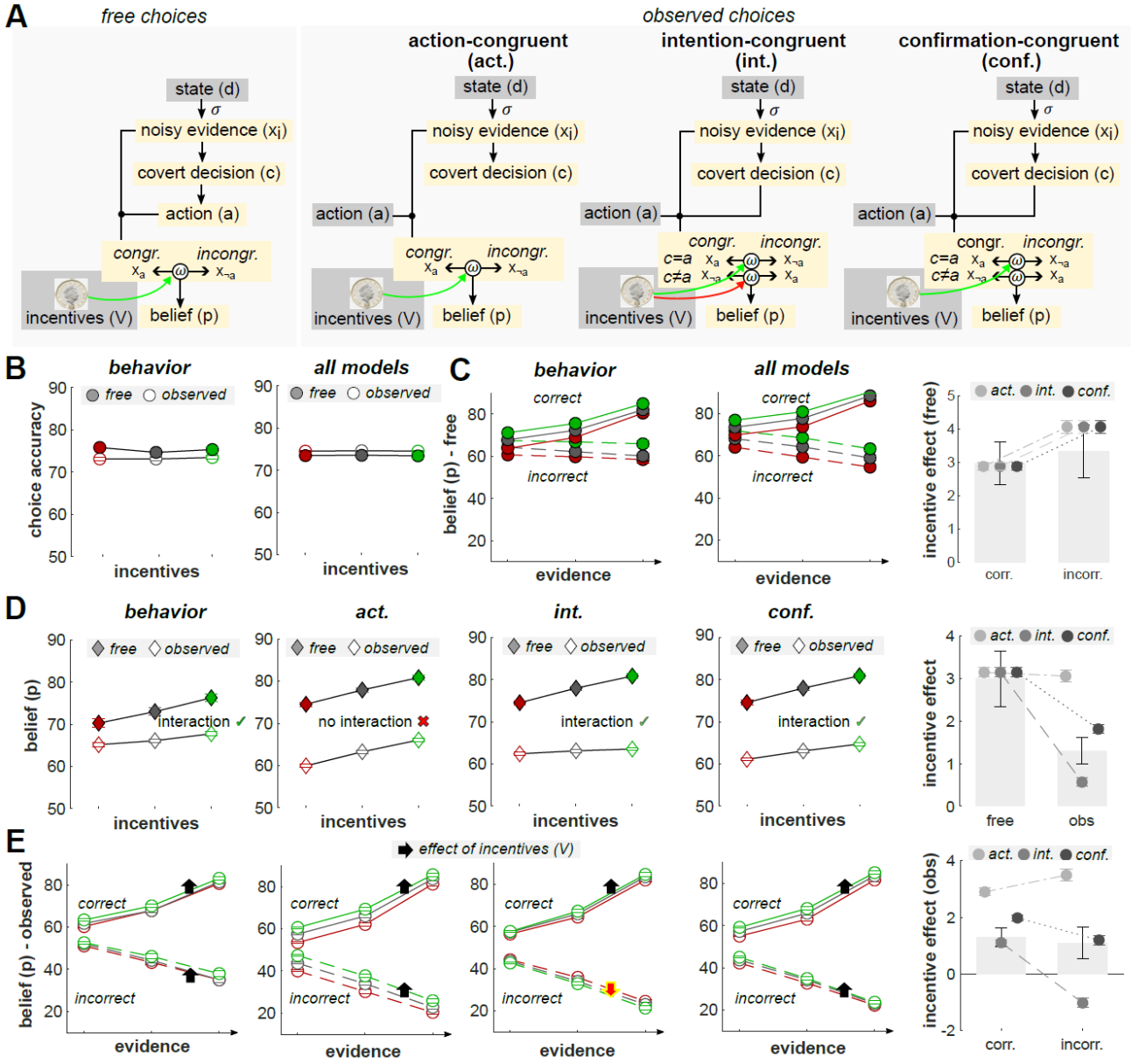

Figure 4 | Models of belief formation and predictions compared to data. **A.** Models of biased beliefs in Free and Observed choices. In Free choices, noisy evidence leads to a covert commitment which translates into action. Belief in the action being correct is based on a weighted difference of evidence, with incentive value increasing the weight of the chosen action and decreasing that of the unchosen action. In observed choices, the incentive effect could exaggerate evidence based on the action, the decision, or their interaction. When reporting belief in a chosen action that mismatches the covert decision (a ≠ c), each model affects this bias differently: either keeping it (action-congruent), inverting it (intention-congruent), or removing it (confirmation-congruent). **B.** Non-

discriminative pattern — choice accuracy. Left: human behavior for accuracy over incentive and agency. Right: model predictions. Incentives have no effect on accuracy in behavior nor models, and all models are equivalent for choice. **C.** Non-discriminative pattern — incentive effect on belief in free choices. Left: human behavior for belief across incentives, stimulus evidence, and choice correctness. Middle: model predictions. Right: average incentive slopes split by correctness, for behavior (bars) and models (circles). In both behavior and simulations, incentives increase belief for stimuli of varying difficulty, both for correct and incorrect choices. In free choices all models are equivalent and predict the same effect, since action always matches covert decision (a = c). **D.** First discriminative pattern — incentive × agency interaction effect on beliefs. In left-right order: human behavior, model simulations (action- intention- and confirmation-congruent model respectively), and average incentive slopes split by agency condition. In behavior, the incentive effect is reduced in observed choices. The action-congruent model does not reproduce this effect, whereas the intention- and confirmation-congruent do. **E.** Second discriminative pattern — positive incentive effect in beliefs for both correct and incorrect observed choices. In left-right order: human behavior (belief in observed decisions across incentives, stimulus evidence, and correctness), model predictions, and average incentive slopes for behavior and models, split by correctness. In behavior, the incentive effect is positive across correct and incorrect choices. This is reproduced by the action-congruent and confirmation-congruent model, but not the intention-congruent model, which predicts a negative incentive effect in incorrect observed choices.

Discussion

In the present study, we investigated the biasing effects of incentive motivation on probabilistic judgments. Our main result points toward a robust effect of the net-incentive value –and not of absolute incentive values– on confidence in self-generated decisions. Individuals are more confident in gain contexts, and less confident in loss contexts. This constitutes a new conceptual replication of our previous studies (Hoven, Brunner, et al., 2022; Hoven, de Boer, et al., 2022; Lebreton et al., 2018, 2019; Salem-Garcia et al., 2023; Ting et al., 2020, 2023) and similar ones (Bröder et al., 2025), generalizing the effect to a new task (RDK), a new experimental setup (0-100% scale) and a new subject pool (Prolific sample). In addition, we repeatedly show that the same biasing effect of incentives is significantly attenuated for first-order beliefs (when judgment is about the correctness of an externally-generated decision), and for confidence judgments about non-immediately-self-generated decisions. The current set of results has important implications.

First, several behavioural patterns refine our understanding of the biasing effects of incentives on confidence judgments. The attenuation of the net-incentive value effect on belief clearly rules out the hypothesis that the effect on confidence could trivially be due to an unspecific Pavlovian effect in which gains and losses simply shift ratings, respectively, up and down. In a way, this result reinforces the idea that the biasing effects of incentive on confidence judgments are a genuine affective and motivational bias. In the appendix of the present manuscript, we also report on analyses made on confidence calibration – i.e. on the direct contrast between the reported p(correct) and the actual decision-accuracy, which provides an estimate of how over- or under-confident/optimistic our participants are (see **Figure S1**). As opposed to our previous studies (Hoven, Brunner, et al., 2022; Hoven, de Boer, et al., 2022; Lebreton et al., 2018, 2019; Salem-Garcia et al., 2023; Ting et al., 2020, 2023), participants did not systematically exhibit overconfidence in the neutral condition. For instance, in Experiments 2-3, participants appear underconfident, yet still exhibit an effect of the net incentive value on calibration. This important result notably demonstrates that, symmetrically to gains, losses do bias confidence downward rather than improve calibration. Finally, the effect of incentive magnitude in Exp.1b reassesses that the bias is driven by the net incentive value – rather than by the incentive valence (Lebreton et al., 2018).

Overall, both results (effects of net incentive value rather than absolute incentive value; differential effect depending on choice agency) initially appear at odds with a general interpretation of motivated beliefs or desirability bias theories. As illustrated in **Figure 2A**, under this hypothesis, one would indeed prefer to believe in a higher p(correct) in both gains and losses conditions in order to contemplate a higher putative payoff. This would predict an effect of absolute incentive value on the reported p(correct) instead of the robust effect of the net incentive value that we have established. Beyond their intuitive appeal and their apparent ubiquity, one should therefore resist the temptation to simplistically and abusively generalize the applicability of desirability biases and motivated beliefs (Logg et al., 2018). Actually, in the present experimental setup, the relative timing of incentive within the trial (after stimulus and decision), their relatively small monetary value and ego-relevance are all factors that may limit the possibility for agent to engage into the kind of belief distortions mechanisms outlined in modern theories of decision-making (Bénabou & Tirole, 2016; Epley & Gilovich, 2016; Krizan & Windschitl, 2007).

Because our elicitation and payment procedure produce similar accuracy incentives and/or belief-distortions incentives in free-choice and observed-choice conditions, it was legitimate to expect monetary incentives to have similar effects in those conditions. As reported throughout this manuscript, we nonetheless consistently observed that incentive effects were attenuated in the observed-choice condition, as well as associated conditions (forced-choice and replayed-choice), compared to the free-choice condition. It is quite remarkable that this bias persists in Exp.2, where participants could, in the observed-choice condition, make a judgment about the observed choice that could still be conceptualized as a combination of a binary choice (with the two halves of the scale) and of a confidence judgment. Consistent with the discrepancy that we observed nonetheless, a couple of studies reported that one stage versus two stage confidence elicitations (respectively corresponding to our observed-choice versus free-choice condition in Exp.2) triggered different cognitive processes and computations, and generated different behavioural signatures (Aitchison et al., 2015; Pereira et al., 2020). In the absence of a unifying model, it would therefore be tempting to infer that incentive motivation affects beliefs and confidence judgment differently, potentially through fundamentally different mechanism. This hypothesis is in line with the fact that, according to some definitions, probabilistic judgments in the free-choice condition (i.e.

confidence judgments; as opposed to the judgments made in the observed-choice condition) constitute metacognitive judgments. Metacognitive operations are known to be dissociable from underlying cognition "first-order" processes, and rely on different sets of computational and neural resources (Fleming & Daw, 2017; Fleming & Dolan, 2012; Pouget et al., 2016), such that their relation with motivational processes could involve specific neuro-computational pathways. Incidentally, our original replayed-choice and forced-choice conditions somewhat blur this distinction between cognitive and metacognitive probabilistic judgments, and might ultimately prove useful to question some specific tenets of theories of metacognition.

At the opposite of the spectrum of possible explanations for this discrepancy between free- and observed choice conditions, we propose a unifying model featuring a generic mechanism of incentive sensitivity for probabilistic judgments, thereby common to all conditions. This model is based on four fundamental principles. First, as is now almost customary in computational cognitive (neuro) science, we assume that our agents' integration of the stimulus information is noisy, which can lead to (confident) misperception (Fleming & Daw, 2017; Sanders et al., 2016). Second, we assume that agents form *covert decisions* – i.e. make commitments to perceptual decisions that predate action– in both free-choice and observed-choice conditions. An increasing number of findings support this idea, leveraging empirical evidence from the decoding of pre-decision neural signals with functional neuroimaging modalities (Balsdon et al., 2021; Charles et al., 2014). Third, we assume that probabilistic judgments build on an asymmetric balance between evidence that is congruent versus incongruent *with the action-confirmed covert decision.* This idea is increasingly popular in the field of confidence judgments (Miyoshi & Lau, 2020; Peters et al., 2017), and could result from a consistency bias in the integration of evidence, akin to a form of covert-decision confirmation bias (Glickman et al., 2022). Our final assumption, more specific to the current experimental setup and set of results, states that monetary incentives modulate the balance between the evidence that is congruent and incongruent with the (action-confirmed and covert) decision. Indeed, intuitively, gain-framing increases the relative focus on the decision being correct ("you can *win* 1 euro *if correct*"), while the loss-framing shift this relative focus on the decision being incorrect ("you can *lose* 1 euro *if incorrect*") – see (Sakamoto & Miyoshi, 2024) for a similar rationale.

This model is actually closely related to one we proposed to account for the emergence of confidence in reinforcement-learning, where the effects of outcome valence (gain versus loss frame) also naturally arises via an overweighting of choice-congruent evidence in the building of confidence (Salem-Garcia et al., 2023; Ting et al., 2023). Interestingly, a recent study also found that, in a similar reinforcement-learning setup, imposing choices (akin to our observed-choice condition) eliminated a learning bias (the asymmetric, confirmatory updating) that can otherwise be mechanistically related to confidence biases (Chambon et al., 2020; Salem-Garcia et al., 2023). These reports are consistent with the idea that confidence biases could be linked to choice-related processes − or even to cognitive processes that precede the actual choice (Samaha & Denison, 2020)−, that agency interacts with metacognition (Charles et al., 2020; Wen et al., 2023), and that the computations underlying confidence judgments depend on the framing of the goal of decision-making tasks (Sepulveda et al., 2020).

Overall, our study sheds new light on the mechanisms governing the formation of beliefs and confidence judgments and their interaction with motivational and affective processes. Our results also highlight the fundamental limitations of incentive-based elicitation mechanisms, which, although theoretically appealing, unavoidably trigger motivational and affective biases that can hamper the desired properties of those "truth-serums". Thereby, we also reveal the critical impact of apparently inconsequential dimensions of probabilistic judgment elicitation, such as incentive framing and one-step vs two-step judgments). Finally, these investigations expose the limits of popular behavioral theories (motivated beliefs, desirability bias), which appear less ubiquitous than sometimes thought, and whose applicability to different contexts should be carefully evaluated.

Methods

Subjects, payment and ethics

A total of 800 participants took part in our online experiments. Participants were excluded and replaced if they failed an attention check which was displayed at the end of the first and fourth block of the main task. This attention check consisted of a screen indicating the participant to press a number from 1 to 5, randomly selected by the computer. If they pressed the wrong key, the check was repeated up to 5 times. The check was considered failed if the participant pressed the wrong key 5 times. Participants were recruited and paid via the Prolific service. The study complied with the ethical rules of Prolific, of the Universite de Lorraine, and more generally, with the Declaration of Helsinki. There was no deception involved. Participants were paid a base show-up fee of £3.5 (£7 for the replayed condition) plus a £2 endowment in experiments with loss conditions, and could additionally win or lose money throughout the task: 5 trials per (non-zero) incentive condition were sampled at random and effectively paid according to the relevant incentivization scheme (see below). On average, participants spent around 40 minutes on our task (60 in the Replayed condition), and obtained a bonus of £4.93 in Experiments 1a,2 and 3, and £4.20 in Experiment 1b.

Participants and trials exclusion

Participants were excluded if they had participated in other of our experiments and/or based on data quality checks (accuracy below 51%, more than 90% of choices in the same direction, or range of ratings below 10% of the scale)

Trials were excluded from the analysis if they exhibited particularly slow response times (RT>20s for choice or rating)

Experiment 1a. Four participants in the Free condition were excluded from analysis based on data quality checks. Across participants, 115 slow trials were excluded.

Experiment 1b. One participant from the Free condition was excluded from the analysis based on data quality checks. Seven more participants in the Free condition were excluded due to having participated in other of our experiments. Across participants, 100 slow trials were excluded.

Experiment 2. Four participants in the Free choice condition were excluded from analysis (three due to quality criteria, one due to having already taken part in the Observed condition). Across participants, 108 slow trials were excluded.

Experiment 3. Three participants were excluded from analysis based on data quality criteria, and one due to having participated in another of our experiments, leaving 98 participants per condition. Across participants we excluded 152 slow trials.

Basic experimental setup

Experimental design – trial structure

Regardless of the experiment and the agency condition (See details below), all trials shared the same structure. Each trial started with a brief fixation step, on a central fixation cross (750ms). Then, a random-dot kinematogram (RDK) stimulus was displayed for 750ms, followed by decision screen, with left- and right- pointing white arrows respectively arranged on the left and right half of the screen. In the free-choice condition (experiments 1 and 2), participants were required to produce a self-paced response (pressing keys 1 or 3 on the keyboard to select, respectively, the left or right pointing arrows) corresponding to what they identified as the dominant dot-motion direction. The keys were visually represented under their corresponding arrows. The pressed key and corresponding selected arrow were then visually highlighted (arrow turned to red, and key darkened; 200ms). In the observed condition (experiments 1 and 2), no key was displayed, and one arrow was selected by the computer and visually highlighted (blue colour). At the bottom of the screen an instruction "press space to continue" invited participants to press the space bar (self-paced) to proceed with the next steps. In the forced choice condition (experiment 3), one arrow was selected by the computer and highlighted (blue colour). Participants had to press the corresponding key (self-paced), which was then visually highlighted (darkened; 200ms). After this choice step, all agency conditions again shared the same structure: the monetary incentive was revealed (1500ms), together with the text "you can win £X". Then, participants were then invited to rate on a scale their subjective estimate of the probability that the choice was correct. In experiments 1 and 3, the scale ranged from 0 to 100% (from left to right), together with the indication 'wrong' and 'correct' at the left and right extremities. In experiment 2, the scale featured an axial symmetry around the middle point (50%) and incrementally reached 100% at

both the left and right extremities of the scale, which again featured the indications 'wrong' and 'correct'.

Experimental design – stimuli

We used an implementation of RDK with position noise (Newsome & Pare, 1988; Scase et al., 1996). Random Dot Kinematograms (RDK), consist of circular white dots of 0.16 deg diameter, presented within a circular aperture (8 degrees) framed by a grey circumference (1 pixel in thickness). These are presented for 23 frames at ~30 frames per second (~750 ms duration). To achieve an approximate dot density of 16.7 dots/deg^2/s as is common for RDK (Roitman & Shadlen, 2002), 14 dots were presented in each frame ($16.7\pi r^2/30$). Each frame, each dot has a probability of being a "signal dot", and otherwise it is a "noise dot" for that frame. Signal dots move in the coherent direction with a speed of 2 deg/s, and noise dots are repositioned at random within the aperture. On each trial, a motion direction (left or right) and motion coherence level are chosen pseudo-randomly. Coherence refers to the probability of a dot moving at 2 deg/s towards the motion direction.

Experimental design – conditions

All our experiments featured two within subject treatments: incentive and difficulty. Regarding the difficulty treatment, we use 5 levels of motion coherence (0.01, 0.03, 0.05, 0.09, 0.38) to target evenly spaced levels of accuracy from 55 to 95%, based on results from a previous pilot (see **Supplementary Online Materials**). This treatment was not mentioned to participants, nor explicitly signalled in the experiment.

Regarding the incentive treatment, we implemented 2 (Experiment 1b) or 3 levels (Experiments 1a, 2 and 3) with respectively two magnitude levels (low: £0.1; and high: £1) or 3 valence levels (loss: £−1; neutral: £0; gain: £+1).

In Experiments 1a, 2 and 3, participants completed 8 blocks of 30 trials each (5 coherences x 3 incentives x 2 motion directions). In Experiment 1b, participants completed 12 blocks of 20 trials each, (5 coherences x 2 incentives x 2 motion directions).

In Experiments 1and 2, participants were randomly assigned to either the free-choice or the observed-choice condition. In experiment 3, participants were randomly assigned to either the forced-choice or the replayed-choice condition. In the observed-choice condition of Experiment 1b, the computer choices were correct with probability = 0.75 with replacement. In Experiments 1a, 2 and 3, observed and forced choices used different

probabilities for each difficulty level, based on the accuracies of the first 50 participants of the free condition in Experiment 2 (probabilities of 0.5504, 0.66, 0.7158, 0.8325 and 0.9692 from lowest to highest coherence). Participants were informed that the computer choices actually matched previous participants' behaviour, and that they, thereby, reached an averaged human-like accuracy of 75%.

Belief incentivization

At the end of each experiment, five trials from each incentive condition are sampled for a Matching Probability auction (Ducharme & Donnell, 1973; Hollard et al., 2016; Massoni et al., 2014), and visually represented. The details are as seen in previous work (Lebreton et al., 2018): in conditions with gains, participants could earn the corresponding amount when winning the auction, whereas in conditions with losses, they could lose the corresponding amount when losing the auction.

Experimental session structure

At the beginning of each experiment, we use a size calibration procedure to determine the on-screen size of stimuli. Participants are shown a shape made to be reminiscent of the outline of a vertically extended thumb. Participants are instructed to adjust the size of the shape so that it is just covered by their thumb held at arm's length, which is known to be ~2 degrees of visual angle (O'Shea, 1991). This lets us approximate the equivalence of screen space to degrees of visual angle.

After size calibration, participants go through a motion direction discrimination practice block, where they are explained the basic setup of the RDK stimuli (two directions with equal probability, variable difficulty across trials, etc). This practice block has12 trials, where participants are shown a fixation cross for 750ms, followed by the RDK stimulus for 750ms, after which they are indicated to make a response (self-paced) with the left and right arrow keys to indicate the direction of the moving dots (left and right respectively). They were then shown verbal feedback ("Correct" or "Wrong") for 1s. In the first 2 trials we use a coherence of 1 for the RDK, and in the following 10 trials we use the levels described in the design. Coherence and motion direction (left or right) are randomized without replacement and equal probabilities.

Participants also underwent 10 rating practice trials which included ratings as in the main task (see below), matching the choice condition assigned. They were explained the choice condition assigned and belief rating and incentivization mechanism. At the end of each trial, the Matching Probability auction (Ducharme & Donnell, 1973; Hollard et al., 2016) is represented as in Lebreton et al (2018), in order to familiarize participants with the reward procedure used in the main task.

Before the main task (or the belief part of the main task in the replayed condition), participants are explained the payoff structure.

In the free, observed, and forced conditions, participants go through discrimination and rating practice, followed by the main task. In the replayed condition, participants first go through discrimination practice, then main task choices, followed by belief rating practice, and finally main task belief ratings.

Behavioural variables and analysis

Our main analyses focus on two behavioural variables: choice accuracy, and belief (reported p(correct)).

Accuracy is an indicator variable that takes the value 1 if the choices is correct (i.e. if the selected direction matches the dominant direction of the RDK) and 0 otherwise. Note that in the observed- and forced- choice conditions, the accuracy correspond to the computer-selected choices, and slightly varies between individuals and conditions due to the sampling with replacement procedure (see the ***Experimental design – conditions*** section for details).

The reported p(correct) corresponds to the probability indicated on the rating scale.

Our primary analyses consist in LME and GLME regressions with formula y ~ V + |V| + (1 + V + |V| |participant) ran, separately for each choice type condition, on trial-level data. For accuracy, we use logistic regression (we specify binomial noise and logit link function). To address the difference in incentive effects between conditions, we also make regressions for free and observed condition together (y ~ V +|V| + V:Free + |V|:Free + (1 + V + |V| | participant)) as well as for forced and replayed (~ V +|V| + V: Replayed + |V|:Replayed + (1 + V + |V| | participant)). We also fit an extension of this model to Experiment 1a and Experiment 3 together including all four agency conditions, with dummy-coded Agency taking Free choice as the reference (y ~ V *Agency + |V| *

|Agency| + (1|V + |V| | participant)). We then compare the marginal incentive slopes between the different choice groups via the *emtrends* function of the *emmeans* R package (Lenth, 2025), using the Holm-Bonferroni correction for multiple comparisons. To analyze calibration (belief – accuracy), we use participant averages per condition (combination of agency and incentive value). We fit this response variable with LME regressions as those specified above for belief, except without random slopes.

These analyses were performed using R (R Core Team, 2025) . The analyses of individual experiments ran R code through the python interface *pymer4* (Jolly, 2018)

Computational models

We propose a computational framework based on 2D Signal Detection Theory to describe biased belief formation in Free and Observed choices.

Each trial, a state d is drawn uniformly from left vs right motion ($d = -1$ or $d = 1$), with stimulus coherence *coh*. These determine the direction and magnitude of the objective signal s = (s_R, s_L) (right and left channels respectively):

$$\boldsymbol{s} = \begin{cases} (coh, 0), & d = 1 \\ (0, coh), & d = -1 \end{cases}$$

An evidence variable x = (x_R, x_L) (right and left channels respectively) is obtained as a combination of the signal with additive noise $\boldsymbol{\varepsilon}$, and lapses L that set evidence to zero:

$$\boldsymbol{x} = (\boldsymbol{s} + \boldsymbol{\varepsilon})(1 - L)$$

Where $\boldsymbol{\varepsilon}$ is independent zero-mean Gaussian noise with standard deviation σ, and lapses (L = 1) occur with probability λ:

$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$$

$$\mathrm{L} \sim \mathrm{Bernoulli}\ (\lambda)$$

The agent then makes a covert commitment c based on this noisy evidence and a bias β:

$$c = sign(x_R - x_L + \beta)(1 - L)$$

When the agent is allowed to choose (i.e in the Free condition), this covert commitment is translated into action a:

$$\begin{cases} a = c\,, & c \neq 0 \\ a = random\ in\ \{-1,1\}, & c = 0 \end{cases}$$

This makes the choice policy equivalent to a cumulative Gaussian psychometric function with lapses:

$$P(a=1) = \frac{\lambda}{2} + (1-\lambda)\Phi(\frac{s_R - s_L + \beta}{\sqrt{2\sigma}})$$

where Φ is the standard cumulative Gaussian function.

For the Observed decision case, on the other hand, the action is made independently of the participant's covert commitment.

At the belief-forming stage, we consider evidence in the space of chosen vs unchosen direction (instead of left vs right). We define a belief evidence variable that represents the difference in evidence supporting the chosen and unchosen option (see **Supplementary Online Materials** for additional details):

$$X_{BEL} = x_a - x_{\neg a}$$

As a difference of gaussians, the distribution of X_{BEL} can be defined by $N(a(s_R - s_L), 2\sigma^2)$. The posterior probability of the choice being correct is proportional to this quantity, following a logistic function (**see Supplementary Online Materials** for the analytical derivations).

$$belief = p(correct|X_{BEL}) = \frac{1}{1 + e^{\left(-\frac{\widehat{\mu}}{\sigma^2}X_{BEL}\right)}}$$

Where $\hat{\mu}$ is the agent's estimation of the average evidence magnitude in the task, which we set to the value of X_{BEL} needed for 75% accuracy given the agent's σ (see **Supplementary Online Material**)

We now consider a set of biased models where the balance of chosen and unchosen evidence is broken by a coefficient w, which is a function of incentives, commitment, and action:

$$w = f(V, c, a)$$

This coefficient respectively exaggerates and attenuates action-congruent and action-incongruent evidence:

$$X_{BEL} = (1 + w)x_a - (1 - w)x_{\neg a}$$

We consider three alternative bias models, which differ in how incentives interact with commitment and action to affect w. More particularly, these models differ in how they behave when covert commitment and action mismatch.

An Action-Congruent bias simply exaggerates evidence for the action in proportion to the incentive value, making w a linear function of incentive value:

$$w = \alpha V$$

An Intention-Congruent bias exaggerates evidence for the covert commitment in proportion to incentive value. This can be expressed as the bias changing sign when the covert commitment and actual choice mismatch:

$$w = \begin{cases} \alpha V, & a = c \\ -\alpha V, & a \neq c \end{cases}$$

Finally, a Confirmation-Congruent model exaggerates evidence supporting the action when it matches the covert commitment but applies no bias otherwise.

$$w = \begin{cases} \alpha \mathrm{V}, & a = c \\ 0, & a \neq c \end{cases}$$

Simulations

We simulated behavior for Free and Observed choices. We used sensory parameters (σ, λ and β) fitted to participants' choices in the Free condition of Experiments 1a and 2 (see **Suppementary Material**), and a constant belief incentive bias α = 0.15. For each model, we simulated behavior for Free and Observed conditions using these parameters (N = 192 per group), varying incentives and coherence as in Exp1a and 2. As in our human experiments, agents in the Observed choice group were given external choices with correct choice probabilities of 0.5504, 0.66, 0.7158, 0.8325 and 0.9692 from lowest to highest coherence. For each agent, we simulated 100 trials per incentive and coherence level (half for each motion direction).

Analysis of model predictions

To analyze model predictions and compare them to our data, we extended our analysis of incentive effects to interactions with choice correctness and stimulus evidence.

We first pooled across experiments to obtain an extended Free choice dataset from the Free conditions of Experiment 1a and Experiment 2 (N = 192), and an extended Observed choice dataset, from the Non-Free conditions (Observed, Forced, and Replayed) of Experiments 1a and Experiment 3 (N = 296).

We fit the following regressions in both data and simulations, individually for each participant:

$$R1: belief \sim incentive$$
$$R2: belief \sim incentive * correct$$
$$R3: belief \sim incentive * correct * evidence$$

For these regressions, incentive was coded numerically for interpretability (-1,0,1), correctness was effect coded (-0.5, +0.5), and evidence was coded as a centered ordinal predictor (-0.5, 0, +0.5) obtained by binning the five coherence levels into low (0.01, 0.03), medium (0.05), and high (0.09, 0.38).

From R2, we obtained the marginal trends conditioned on correct and incorrect choices respectively (using the function *emtrends* from the R package *emmeans*). Similarly, from R3, we obtained the marginal trends incentive:evidence interaction conditioned on correct and incorrect choices respectively.

When fitting R3 to data, the incentive:evidence interaction for incorrect decisions could not be estimated for 7 participants in the Free condition and 1 participant in the Observed condition, due to a lack of incorrect responses for easier trials.

Data and code availability statement

All code and data will be made publicly available, without restrictions, upon acceptance of the manuscript for publication.

References

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly bayesian analysis of confidence in perceptual decision-making. *PLoS Comput Biol*, *11*(10), e1004519.

Babad, E., & Katz, Y. (1991). Wishful Thinking—Against All Odds. *Journal of Applied Social Psychology*, *21*(23), 1921–1938. https://doi.org/10.1111/j.1559-1816.1991.tb00514.x

Balsdon, T., Mamassian, P., & Wyart, V. (2021). Separable neural signatures of confidence during perceptual decisions. *eLife*, *10*, e68491. https://doi.org/10.7554/eLife.68491

Bénabou, R., & Tirole, J. (2002). Self-Confidence and Personal Motivation*. *The Quarterly Journal of Economics*, *117*(3), 871–915. https://doi.org/10.1162/003355302760193913

Bénabou, R., & Tirole, J. (2016). Mindful Economics: The Production, Consumption, and Value of Beliefs. *Journal of Economic Perspectives*, *30*(3), 141–164. https://doi.org/10.1257/jep.30.3.141

Bröder, A., Navarro-Báez, S., & Undorf, M. (2025). Reducing cheap talk? How monetary incentives affect the accuracy of metamemory judgments. *Memory & Cognition*. https://doi.org/10.3758/s13421-024-01679-5

Chambon, V., Théro, H., Vidal, M., Vandendriessche, H., Haggard, P., & Palminteri, S. (2020). Information about action outcomes differentially affects learning from self-determined versus imposed choices. *Nature Human Behaviour*, *4*(10), Article 10. https://doi.org/10.1038/s41562-020-0919-5

Charles, L., Chardin, C., & Haggard, P. (2020). Evidence for metacognitive bias in perception of voluntary action. *Cognition*, *194*, 104041. https://doi.org/10.1016/j.cognition.2019.104041

Charles, L., King, J.-R., & Dehaene, S. (2014). Decoding the Dynamics of Action, Intention, and Error Detection for Conscious and Subliminal Stimuli. *Journal of Neuroscience*, *34*(4), 1158–1170. https://doi.org/10.1523/JNEUROSCI.2465-13.2014

De Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. In *Annales de l'Institut Poincaré* (Vol. 7). Gauthier-Villars.

Ducharme, W. M., & Donnell, M. L. (1973). Intrasubject comparison of four response modes for "subjective probability" assessment. *Organizational Behavior and Human Performance*, *10*(1), 108–117. https://doi.org/10.1016/0030-5073(73)90007-X

Eil, D., & Rao, J. M. (2011). The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–138. https://doi.org/10.1257/mic.3.2.114

Epley, N., & Gilovich, T. (2016). The Mechanics of Motivated Reasoning. *Journal of Economic Perspectives*, *30*(3), 133–140. https://doi.org/10.1257/jep.30.3.133

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(4), 552–564. https://doi.org/10.1037/0096-1523.3.4.552

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114. https://doi.org/10.1037/rev0000045

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Phil. Trans. R. Soc. B*, *367*(1594), 1338–1349. https://doi.org/10.1098/rstb.2011.0417

Giardini, F., Coricelli, G., Joffily, M., & Sirigu, A. (2008). Overconfidence in Predictions as an Effect of Desirability Bias. In P. M. Abdellaoui & P. D. J. D. Hey (Eds.), *Advances in Decision Making Under Risk and Uncertainty* (pp. 163–180). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-68437-4_11

Glickman, M., Moran, R., & Usher, M. (2022). Evidence integration and decision confidence are modulated by stimulus consistency. *Nature Human Behaviour*, *6*(7), Article 7. https://doi.org/10.1038/s41562-022-01318-6

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Sciences*, *1*(2), 78–82. https://doi.org/10.1016/S1364-6613(97)01014-0

Hollard, G., Massoni, S., & Vergnaud, J.-C. (2016). In search of good probability assessors: An experimental comparison of elicitation rules for confidence judgments. *Theory and Decision*, *80*(3), 363–387. https://doi.org/10.1007/s11238-015-9509-9

Holt, C. A., & Smith, A. M. (2009). An update on Bayesian updating. *Journal of Economic Behavior & Organization*, *69*(2), 125–134. https://doi.org/10.1016/j.jebo.2007.08.013

Hoven, M., Brunner, G., de Boer, N. S., Goudriaan, A. E., Denys, D., van Holst, R. J., Luigjes, J., & Lebreton, M. (2022). Motivational signals disrupt metacognitive signals in the human ventromedial prefrontal cortex. *Communications Biology*, *5*(1), Article 1. https://doi.org/10.1038/s42003-022-03197-z

Hoven, M., de Boer, N. S., Goudriaan, A. E., Denys, D., Lebreton, M., van Holst, R. J., & Luigjes, J. (2022). Metacognition and the effect of incentive motivation in two compulsive disorders: Gambling disorder and obsessive–compulsive disorder. *Psychiatry and Clinical Neurosciences*, *76*(9), 437–449. https://doi.org/10.1111/pcn.13434

Jeffrey, R. C. (1990). *The logic of decision*. University of Chicago press.

Jolly, E. (2018). Pymer4: Connecting R and Python for Linear Mixed Modeling. *Journal of Open Source Software*, *3*(31), 862. https://doi.org/10.21105/joss.00862

Karni, E. (2009). A Mechanism for Eliciting Probabilities. *Econometrica*, *77*(2), 603–606. https://doi.org/10.3982/ECTA7833

Koellinger, P., & Treffers, T. (2015). Joy Leads to Overconfidence, and a Simple Countermeasure. *PLOS ONE*, *10*(12), e0143263. https://doi.org/10.1371/journal.pone.0143263

Krizan, Z., & Windschitl, P. D. (2007). The influence of outcome desirability on optimism. *Psychological Bulletin*, *133*(1), 95. https://doi.org/10.1037/0033-2909.133.1.95

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480.

Lebreton, M., Bacily, K., Palminteri, S., & Engelmann, J. B. (2019). Contextual influence on confidence judgments in human reinforcement learning. *PLOS Computational Biology*, *15*(4), e1006973. https://doi.org/10.1371/journal.pcbi.1006973

Lebreton, M., Langdon, S., Slieker, M. J., Nooitgedacht, J. S., Goudriaan, A. E., Denys, D., van Holst, R. J., & Luigjes, J. (2018). Two sides of the same coin: Monetary incentives concurrently improve and bias confidence judgments. *Science Advances*, *4*(5). https://doi.org/10.1126/sciadv.aaq0668

Lenth, R. V. (2025). *emmeans: Estimated Marginal Means, aka Least-Squares Means* [Computer software]. https://rvlenth.github.io/emmeans/

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases* (pp. 306–334). Cambridge University Press. http://www.cambridge.org/emea/

Loewenstein, G., & Molnar, A. (2018). The renaissance of belief-based utility in economics. *Nature Human Behaviour*, *2*(3), 166–167. https://doi.org/10.1038/s41562-018-0301-z

Logg, J. M., Haran, U., & Moore, D. A. (2018). Is overconfidence a motivated bias? Experimental evidence. *Journal of Experimental Psychology: General*, *147*(10), 1445–1465. https://doi.org/10.1037/xge0000500

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

Manski, C. F. (2004). Measuring Expectations. *Econometrica*, *72*(5), 1329–1376. https://doi.org/10.1111/j.1468-0262.2004.00537.x

Massoni, S. (2014). Emotion as a boost to metacognition: How worry enhances the quality of confidence. *Consciousness and Cognition*, *29*, 189–198. https://doi.org/10.1016/j.concog.2014.08.006

Massoni, S., Gajdos, T., & Vergnaud, J.-C. (2014). Confidence measurement in the light of signal detection theory. *Frontiers in Psychology*, *5*, 1455. https://doi.org/10.3389/fpsyg.2014.01455

Miyoshi, K., & Lau, H. (2020). A decision-congruent heuristic gives superior metacognitive sensitivity under realistic variance assumptions. *Psychological Review*, *127*(5), 655–671. https://doi.org/10.1037/rev0000184

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502.

Newsome, W. T., & Pare, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *The Journal of Neuroscience*, *8*(6), 2201–2211. http://www.ncbi.nlm.nih.gov/pubmed/3385495

O'Shea, R. P. (1991). Thumb's rule tested: Visual angle of thumb's width is about 2 deg. *Perception*, *20*(3), 415–418. https://doi.org/10.1068/p200415

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

Pereira, M., Faivre, N., Iturrate, I., Wirthlin, M., Serafini, L., Martin, S., Desvachez, A., Blanke, O., Van De Ville, D., & Millán, J. del R. (2020). Disentangling the origins of confidence in speeded perceptual judgments through multimodal imaging. *Proceedings of the National Academy of Sciences*, *117*(15), 8382–8390. https://doi.org/10.1073/pnas.1918335117

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

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374. https://doi.org/10.1038/nn.4240

R Core Team. (2025). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ramsey, F. (1931). Truth and Probability. In *The Foundations of Mathematics and other Logical Essays: Vol. Ch. VII* (R.B. Braithwaite, pp. 156–198). Kegan, Paul, Trench, Trubner & Co.

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *22*(21), 9475–9489.

Sakamoto, Y., & Miyoshi, K. (2024). A confidence framing effect: Flexible use of evidence in metacognitive monitoring. *Consciousness and Cognition*, *118*, 103636. https://doi.org/10.1016/j.concog.2024.103636

Salem-Garcia, N., Palminteri, S., & Lebreton, M. (2023). Linking confidence biases to reinforcement-learning processes. *Psychological Review*, *130*(4), 1017–1043. https://doi.org/10.1037/rev0000424

Samaha, J., & Denison, R. (2020). *The positive evidence bias in perceptual confidence is not post-decisional* [Preprint]. Animal Behavior and Cognition. https://doi.org/10.1101/2020.03.15.991513

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron*, *90*(3), 499–506. https://doi.org/10.1016/j.neuron.2016.03.025

Savage, L. J. (1954). *The Foundations of Statistics*. Courier Dover Publications.

Scase, M. O., Braddick, O. J., & Raymond, J. E. (1996). What is Noise for the Motion System? *Vision Research*, *36*(16), 2579–2586. https://doi.org/10.1016/0042-6989(95)00325-8

Schlag, K. H., Tremewan, J., & van der Weele, J. J. (2015). A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, *18*(3), 457–490. https://doi.org/10.1007/s10683-014-9416-x

Schotter, A., & Trevino, I. (2014). Belief elicitation in the laboratory. *Annu. Rev. Econ.*, *6*(1), 103–128.

Sepulveda, P., Usher, M., Davies, N., Benson, A. A., Ortoleva, P., & De Martino, B. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *eLife*, *9*, e60705. https://doi.org/10.7554/eLife.60705

Sharot, T. (2011). The optimism bias. *Current Biology*, *21*(23), R941–R945. https://doi.org/10.1016/j.cub.2011.10.030

Shekhar, M., & Rahnev, D. (2021). Sources of Metacognitive Inefficiency. *Trends in Cognitive Sciences*, *25*(1), 12–23. https://doi.org/10.1016/j.tics.2020.10.007

Siedlecka, M., Koculak, M., & Paulewicz, B. (2021). Confidence in action: Differences between perceived accuracy of decision and motor response. *Psychonomic Bulletin & Review*, *28*(5), 1698–1706. https://doi.org/10.3758/s13423-021-01913-0

Ting, C.-C., Palminteri, S., Engelmann, J. B., & Lebreton, M. (2020). Robust valence-induced biases on motor response and confidence in human reinforcement learning. *Cognitive, Affective, & Behavioral Neuroscience*. https://doi.org/10.3758/s13415-020-00826-0

Ting, C.-C., Salem-Garcia, N., Palminteri, S., Engelmann, J. B., & Lebreton, M. (2023). Neural and computational underpinnings of biased confidence in human reinforcement learning. *Nature Communications*, *14*(1), 6896. https://doi.org/10.1038/s41467-023-42589-5

Van den Steen, E. (2004). Rational Overoptimism (and Other Biases). *American Economic Review*, *94*(4), 1141–1151. https://doi.org/10.1257/0002828042002697

Wallsten, T. S., & Budescu, D. V. (1983). State of the Art—Encoding Subjective Probabilities: A Psychological and Psychometric Review. *Management Science*, *29*(2), 151–173. https://doi.org/10.1287/mnsc.29.2.151

Wen, W., Charles, L., & Haggard, P. (2023). Metacognition and sense of agency. *Cognition*, *241*, 105622. https://doi.org/10.1016/j.cognition.2023.105622

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321. https://doi.org/10.1098/rstb.2011.0416

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