How Overconfidence Bias Influences Suboptimality in Perceptual Decision Making

Marine Hainguerlot, Thibault Gajdos, Jean-Christophe Vergnaud, Vincent de Gardelle

How overconfidence bias influences suboptimality in perceptual decision making

Marine Hainguerlot[1], Thibault Gajdos[2], Jean-Christophe Vergnaud[3¶] and Vincent de Gardelle[4¶]

[1]Erasmus School of Economics, Erasmus University Rotterdam, the Netherlands.

[2]Aix Marseille University, CNRS, LPC, Marseille, France

[3]Centre d'Economie de la Sorbonne, CNRS UMR 8174, Paris, France.

[4]CNRS and Paris School of Economics, Paris, France.

[¶] denotes equal contribution

Word count: 6,660

Author Note

This research was supported by the Agence Nationale de la Recherche (ANR-16-CE28-0002, ANR-16-ASTR-0014 to V.d.G.), the Région Provence-Alpes-Côte d'Azur (CREMCO, 2012-2016 to T.G.), and by University Paris 1 (doctoral grant to M.H.). The authors declare no competing interests.

M.H., T.G., J.C.V., and V.d.G. developed the theory and designed the experiment. M.H. conducted the experiment. M.H., J.C.V., and V.d.G. analyzed the results. M.H. and V.d.G. prepared the figures. M.H., J.C.V., and V.d.G. wrote the manuscript.

All data, analysis codes, and research materials are available on the Open Science Framework repository (https://osf.io/4qw9e/?view_only=48bae1de632c4ff895cfa49743b41dfa)

Corresponding author: Marine Hainguerlot, P.O. Box 1738 Rotterdam, 3000 DR, The Netherlands.

hainguerlot@ese.eur.nl

Abstract

In perceptual decision making, it is often found that human observers combine sensory information and prior knowledge suboptimally. Typically, in detection tasks, when an alternative is a priori more likely to occur, observers choose it more frequently to account for the unequal base rate but not to the extent they should, a phenomenon referred to as "conservative decision bias" (i.e., observers do not shift their decision criterion enough). One theoretical explanation of this phenomenon is that observers are overconfident in their ability to interpret sensory information, resulting in overweighting the sensory information relative to the prior knowledge. Here, we derived formally this candidate model and we tested it in a visual discrimination task in which we manipulated the prior probabilities of occurrence of the stimuli. We measured confidence in decisions and decision criterion placement in two separate experimental sessions for the same participants (N=69). Both overconfidence bias and conservative decision bias were found in our data, but critically the link that was predicted between these two quantities was absent. Our data suggested instead that when informed about the a priori probability, overconfident participants put less effort into processing sensory information. These findings offer new perspectives on the role of overconfidence bias to explain suboptimal decisions.

Keywords: Overconfidence bias, Perceptual decision making, Suboptimality, Signal Detection Theory, Conservative decision bias, Sensitivity.

Public significance statement: In detection tasks, humans are presented with evidence and must decide whether a target is present or not, for example, whether there are dangerous items on x-ray images of luggage. When their prior knowledge indicates that the target is likely to be present, they should adjust their decision criterion such that they would require less evidence to detect the target. This study highlights that how well people adjust their decision criterion does not depend on their confidence in their ability to interpret the evidence. The results suggest that people who are overconfident in their ability invest less effort in the task when prior knowledge is available.

1. Introduction

Whether human observers can combine optimally multiple pieces of information has been studied across modalities (e.g., Ernst & Banks, 2002), over time (e.g., Yang & Shadlen, 2007), and between individuals (e.g., Bahrami et al., 2010). In perceptual decision-making, it is often found that observers combine sensory information and prior knowledge suboptimally (Rahnev & Denison, 2018). Typically, in detection tasks, when an alternative is a priori more likely to occur, observers choose it more frequently to account for the unequal base rate but not to the extent they should. This phenomenon referred to as "conservative decision bias" (i.e., observers do not shift their decision criterion enough) has been observed with laboratory tasks using basic visual decisions (Ackermann & Landy, 2015; Green & Swets, 1966; Murrell, 1977; Ulehla, 1966) but also in experiments emulating real-world decisions such as the detection of faulty products (Botzer et al.,2010; Botzer et al., 2013; Chi & Drury, 1998), or enemy targets (Wang et al., 2009).

To achieve Bayes- optimal combination, observers must combine the sensory information and the prior knowledge according to its weight of evidence. For instance, when a medical doctor inspecting a chest CT scan must decide whether a lung nodule is cancerous or not, her decision should weigh this piece of sensory information and her prior knowledge of the risk factors for lung cancer accordingly. However, observers often behave as if they under-weigh prior knowledge such that, for example, the doctor would be less reluctant to decide that a lung nodule is cancerous among smokers but not to the extent she should. Critically, in our example, there is a fundamental difference between the risk factors for which the weight of evidence is quantified objectively by epidemiologists and the chest CT scan whose weight of evidence depends on the doctor. If the doctor overestimates her ability to distinguish between cancerous and benign lung nodules (e.g., "I am sure that this nodule has a small size."), she may rely less on the information provided by the risk factor (e.g., "10 % of smokers develop a lung cancer") than what is optimal.

In this paper, we ask whether conservative decision bias is caused by overestimation of one's own ability to process sensory information. Several studies have found that observers overestimate

the accuracy of their decisions (i.e., they are overconfident) in signal detection tasks including in the perceptual domain (Baranski & Petrusic, 1994; Kvidera & Koutstaal, 2008; Massoni et al., 2014). Here, we consider that overconfidence bias captures observers' misestimation of their ability to process sensory information. Such a link between conservative decision bias and misestimation of the sensory information has already been proposed (Kubovy, 1977; Ackermann & Landy, 2015) but it has not been supported by direct empirical evidence. We test this prediction experimentally in a perceptual task.

The hypothesized link between overconfidence bias and conservative decision bias can be formally described within Signal Detection Theory (SDT, Green and Swets, 1966), which provides a framework for analyzing decisions between two options, and for distinguishing between the sensitivity of the observer to the sensory information (d'), and the decision criterion of the observer (c) measuring the amount of sensory evidence for which she is indifferent between the two choice options. Theoretically, this criterion should be affected by information about the a priori probability of occurrence of the choice options. Under this model, we can predict that if an observer is overconfident (i.e., overestimates her own sensitivity), she would also not adjust her response criterion optimally. Formally, as detailed in the Methods, we should have:

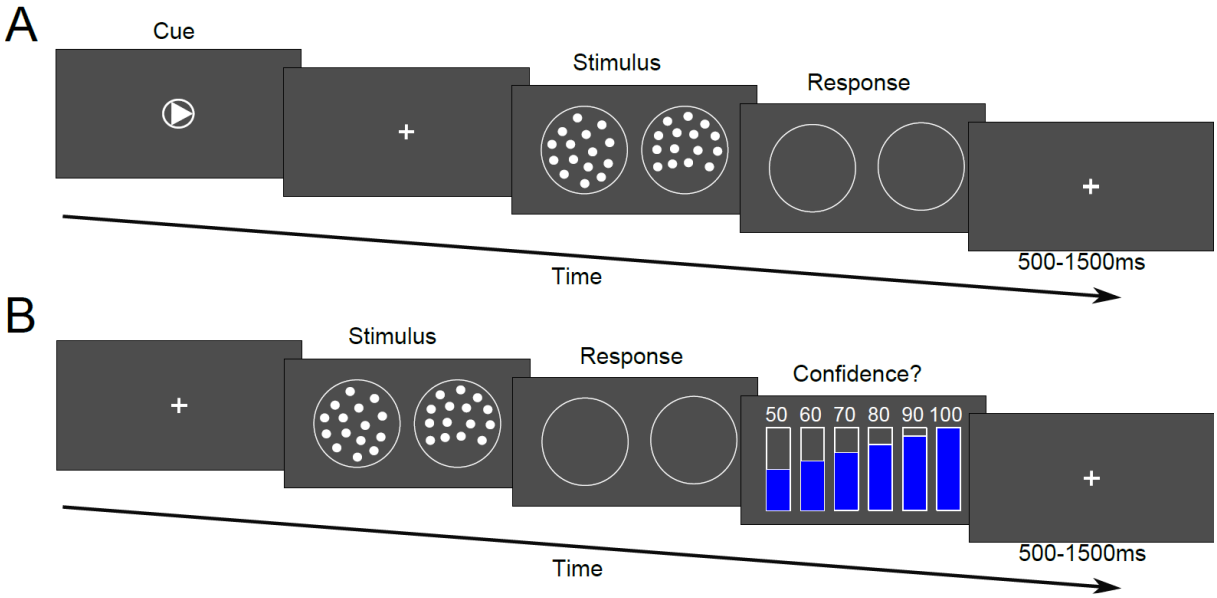
$$c_{subj} = c_{ideal}\frac{d'}{d'_{subj}},$$

where d' and d'_{subj} denote the observer's actual and perceived sensitivity, while c_{ideal} denotes the ideal decision criterion set by an observer perfectly aware of her own sensitivity d', and c_{subj} denotes the criterion that would be used by an observer relying on d'_{subj} instead. Intuitively, the equation means that the criterion set by the observer should deviate from the ideal criterion by a factor that is the inverse of her overconfidence bias, leading to a conservative decision bias if she is overconfident (i.e., $\frac{c_{subj}}{c_{ideal}} < 1 \; if \; d'_{subj} > d'$).

We test this prediction experimentally in a visual discrimination task by measuring in the same participants, but in distinct sessions, both 1) confidence about decisions when the base rate is

equal and 2) criterion adjustments in response to unequal base rates. Briefly, our participants (N=69) had to identify which of two sets presented on the computer screen contained more dots (see Fig 1) in two experimental sessions conducted four days apart. Here, the dots are visible enough but it is hard to identify which set has more dots given the small difference in number of dots between the two sets, and the short presentation time. In the confidence session, after each decision participants indicated their confidence on a quantitative scale. In the cueing session, prior probability about the forthcoming stimulus was given in the form of a symbolic cue. On each trial, the stimulus was preceded by a symbolic cue indicating the correct side with 75% validity or by a neutral cue indicating that both sides were equally likely. Before the task, the meaning and validity of these symbolic cues were fully explained to participants, who were instructed to optimally use both the cue and the stimulus information to maximize their payoff. We then tested whether overconfidence bias measured in the confidence session would relate to conservative decision bias measured in the cueing session, as predicted by our model.

Figure 1. *Experimental paradigm*



Note. Participants had to indicate which circle (left or right) contained more dots. (A) In the cueing session, stimuli were presented after a neutral or 75% valid cue that participants had to optimally use to make their decisions. (B) In the confidence session, decisions were followed by confidence judgments on an incentivized probability rating scale.

2. Methods

2.1. *Transparency and openness*

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study, and we follow JARS (Kazak, 2018). All data, analysis codes, and research materials are available on the Open Science Framework repository (https://osf.io/4qw9e/?view_only=48bae1de632c4ff895cfa49743b41dfa). The experiment was programmed using Psychotoolbox (Brainard, 1997) and MATLAB 8.3 (The Math Works, Inc., 2014). Data were analyzed using MATLAB 8.3 (The Math Works, Inc., 2014), R package lme4 (Bates et al., 2015, R Core Team, 2018), and JASP (JASP Team, 2022). This study's design and its analysis were not pre-registered.

2.2. *Experiment*

2.2.1. A priori power analysis and participants

To test the hypothesized link between overconfidence bias and conservative decision bias, our empirical strategy relied on evaluating the correlation between predicted and actual criteria across participants (for more details about our strategy, see section 2.3.2). Theoretically, if the hypothesized link holds true, the predicted and actual criteria should perfectly positively correlate. However, empirically, we aimed to detect an observable correlation of 0.3. A priori power analysis performed with the GPower software (Faul et al., 2007) indicated that the total sample size required to detect a one-tailed correlation of Cohen's medium effect size r=0.3 between two normally distributed variables, given a significance level α =0.05, and a statistical power level $1-\beta$ =0,80 is N=67.

In total, 69 individuals (39 females; mean age = 23 years, SD = 2.5 years) were recruited through the Laboratory of Experimental Economics of Paris. Participants received 13 Euros for participating plus an incentivized bonus described below. The data was collected in 2014.

2.2.2. Ethic statement

The study was conducted in line with the principles of the Declaration of Helsinki. Written informed consent was obtained from all participants before the experiment. No nominative/identifying information was collected. No health information was collected from participants other than gender and age. The research involved negligible risks. In this situation, as per current French regulations, ethics approval was not required, therefore no IRB was consulted before conducting the study.

2.2.3. Stimuli and Task

Rather than collecting confidence judgments following each decision made in the presence of cues, we decided to collect independent measures to avoid two potential issues: i) that asking explicitly participants to evaluate and verbalize their confidence in their decision might change how they would take into account the cues in their decision process; and, ii) that the manipulation of prior probabilities of occurrence of the stimuli (given in the form of a symbolic cue) might alter how participants would evaluate their confidence in their decision. This would also allow us to compare more directly our data with previous literature studying confidence or cueing. The confidence and cueing sessions took place 4 days apart and their order of presentation was counterbalanced across participants. The experiment was run on screens (resolution 1024 x 768) viewed at normal distance (about 60 cm).

Importantly, in both sessions, participants were asked to perform the same perceptual task with the same type of stimuli. On each trial, after a 250ms fixation cross, two sets of about 100 dots were simultaneously presented for 700ms, one on the left side and one on the right side of the computer screen. Participants had to indicate which set contained more dots, by pressing the corresponding arrow (left or right arrow keys) on the keyboard. After the response, the inter-trial interval was jittered between 0.5s and 1.5s. Participants received no feedback about the accuracy of their decision. Response times shorter than 200ms or longer than 2200ms (from stimulus onset) were discouraged by presenting a "too fast" or "too slow" message on the screen.

2.2.4. Calibration

At the beginning of each session, and for each participant, stimulus difficulty x (i.e., the difference in number of dots between the two circles) was calibrated using a 2-down-1 up rule (Levitt, 1971) to obtain 71% of 'left' or 'right' responses. Specifically, one circle contained 100 dots while the other circle (the stimulus) contained 100 + x dots, and x decreased by one step size after two consecutive correct responses and increased by one step size after one failure. In order to obtain more precise estimates more rapidly, the step size was reduced from 20 (an easy initial value) to 16, 8, 4 and 2 dots on trials 12, 24, 60 and 80 respectively. In addition, to account for the possibility that participants may be biased towards responding more 'left' or 'right', we used two independent and interleaved staircases of 150 trials each, one adjusting the value for the left stimulus (x_l) and one for the right stimulus (x_r). With these data, we estimated, for each participant, a psychometric curve representing the proportion of 'left' responses as a function of the difference in number of dots between the left and right circles, fitted with a cumulative normal distribution. To obtain x_l and x_r, we took the difference in number of dots for which the psychometric curve predicted a 70% and 30% of 'left' responses, rounded to its nearest integer and converted to its absolute value. We then kept these values constant across the session.

2.2.5. Symbolic cueing session

In this session, each trial started with a central cue presented for 250ms, before the fixation cross. The cue was either a triangle pointing to the left or the right side of the screen, indicating the correct response with 75% validity (cue condition), or a diamond providing no information (neutral condition). In the cue-condition, 192 trials (96 left cue, 96 right cue) indicated the correct response (75% valid) and 64 trials (32 left cue, 32 right cue) indicated the wrong response (25% invalid). In the neutral condition, the diamond cue was followed by right and left correct responses equally often (128 left, 128 right). Similarly to Rahnev et al. (2011), the trials in the cue and neutral conditions

were administered in 64 randomly interleaved mini-blocks of 8 trials, with each mini-block including either predictive cues only (left and right, randomly interleaved across trials), or only neutral cues. Each mini-block began with a 1 second presentation of the cue(s) used in the forthcoming eight trials to remind participants. At the beginning of the session, participants were fully informed of the meaning of these cues and of the structure of the blocks. They were instructed to use both the stimulus and the cue to make the best possible decisions. Response accuracy was incentivized: participants won 1 point if correct and lost 1 point if incorrect, and points were converted to a bonus payment at the end of the experiment, with 1 point= 0.02 Euros. A training phase with feedback (96 trials) was included.

2.2.6. Confidence session

In the confidence session, both sides were a priori equally likely to occur and no cue was presented. Each response was followed by a confidence rating, in which participants indicated their subjective belief that their response just given was correct, on a 6 steps scale ranging from 50% confident (i.e., guess) to 100% confident, in steps of 10%. Participants responded using the numerical keys on the top-left of the keyboard. This confidence rating was incentivized using a probability matching rule (Massoni et al., 2014), which is a variant of the Becker-DeGroot-Marschak rule (Becker et al., 1964) classically used in experimental economics. The participant is offered an exchange between his response and a lottery ticket with a probability P of success. The number P is randomly determined on each trial (with a uniform distribution between 0 and 1) and compared to the confidence response. If P is greater than the confidence, then the participant's reward is determined by the lottery. If not, it is determined by the accuracy of the response. The mechanism was presented to participants as a way to maximize their earnings by providing accurate confidence ratings. Instructions, examples, and a training phase with feedback (40 trials) were included to make sure that participants understood the mechanism. Participants then completed 512 trials in the session.

2.2.7. Running memory span task

Both the cueing and confidence sessions started with a standard running memory span task (Pollack et al., 1959; Conway et al., 2005; Broadway & Engle, 2010). Details for this task (which will only be used as a control measure in our final analyses) are presented in the Supplementary Materials (Methods S1).

2.3. *Computational approach*

2.3.1. The SDT model linking overconfidence bias and conservative decision bias

We shall now describe the relation between overconfidence bias and conservative decision bias that we should expect under Signal Detection Theory (SDT), illustrated in Fig 2. Following SDT, let's assume that a given state of nature (i.e., A = the right circle has more dots vs. B = the left circle has more dots) generates an internal sensory signal (denoted x) that the observer compares with a decision criterion (denoted c). The observer responds that the state of nature is "A" when the internal sensory signal is above c. Furthermore, across trials, both states generate normally distributed values of x, with equal variance σ^2 and with means μ_A and μ_B. Without loss of generality and for simplicity, we assume that μ_A and μ_B are symmetric around 0, and that $\sigma = 1$. Defining the observer's sensitivity $d' = |\mu_A - \mu_B|$, the probability distribution of x, given the true state of the nature (either *A* or *B*) is thus given by:

$$P(x|A) = N(x, +d'/2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-d'/2)^2} \quad \text{(eq. 1a)}$$

$$P(x|B) = N(x, -d'/2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x+d'/2)^2}. \quad \text{(eq. 1b)}$$

The logarithm of the likelihood ratio of the sensory signal is then:

$$LS(x) = \log\left(\frac{P(x|A)}{P(x|B)}\right) = -\frac{1}{2}\left(x - \frac{d'}{2}\right)^2 + \frac{1}{2}\left(x + \frac{d'}{2}\right)^2 = d'x \quad \text{(eq. 2)}$$

Assume also that the observer has access to some information about the a priori probability of occurrence of *A* and *B.* To maximize expected accuracy, the observer should set the criterion such

that she responds “A” when the posterior probability that A is present is greater than the posterior probability that B is present. By Bayes’ s rule, this decision rule can be written as a function of the log-odds prior (LP) and the log likelihood ratio of the sensory signal ($LS(x)$). The ideal observer who perfectly estimates her own sensitivity responds according to the sign of the decision variable $DV\ (x)$ defined as follows:

Say "A" $when$:

$$DV(x) = \log\left(\frac{P(A|x)}{P(B|x)}\right) = \log\left(\frac{P(A)}{P(B)}\right) + \log\left(\frac{P(x|A)}{P(x|B)}\right) = LP + LS(x) = LP + \ d'x > 0 \qquad \text{(eq. 3a)}$$

On the other hand, the non-ideal observer who misestimates her own abilities use a subjective value d'_{subj} instead of d' to evaluate the sensory signal, leading to the likelihood ratio $LS_{subj}(x) = \log\left(\frac{P_{subj}(x|A)}{P_{subj}(x|B)}\right)$, where P_{subj} denotes the observer’s subjective probability of observing an internal signal x. In such a case, the non-ideal observer responds according to the sign of the following decision variable:

Say "A" $when$:

$$DV_{subj}(x) = \log\left(\frac{P_{subj}(A|x)}{P_{subj}(B|x)}\right) = LP + LS_{subj}(x) = LP + \frac{d'_{subj}}{d'}(d'x) > 0 \qquad \text{(eq. 3b)}$$

It follows from equation 3b that if the observer is overconfident ($\frac{d'_{subj}}{d'} > 1$) she overweighs the sensory information, whereas if she is underconfident ($\frac{d'_{subj}}{d'} < 1$) she underweights it.

Using equations 3a and 3b, the optimal decision criterion set by the ideal observer and the non-ideal observer are defined respectively as:

$$DV(c_{ideal}) = 0 \Leftrightarrow LP + \ d'c_{ideal} = 0 \Leftrightarrow c_{ideal} = -\frac{1}{d'}LP \qquad \text{(eq.4a)}$$

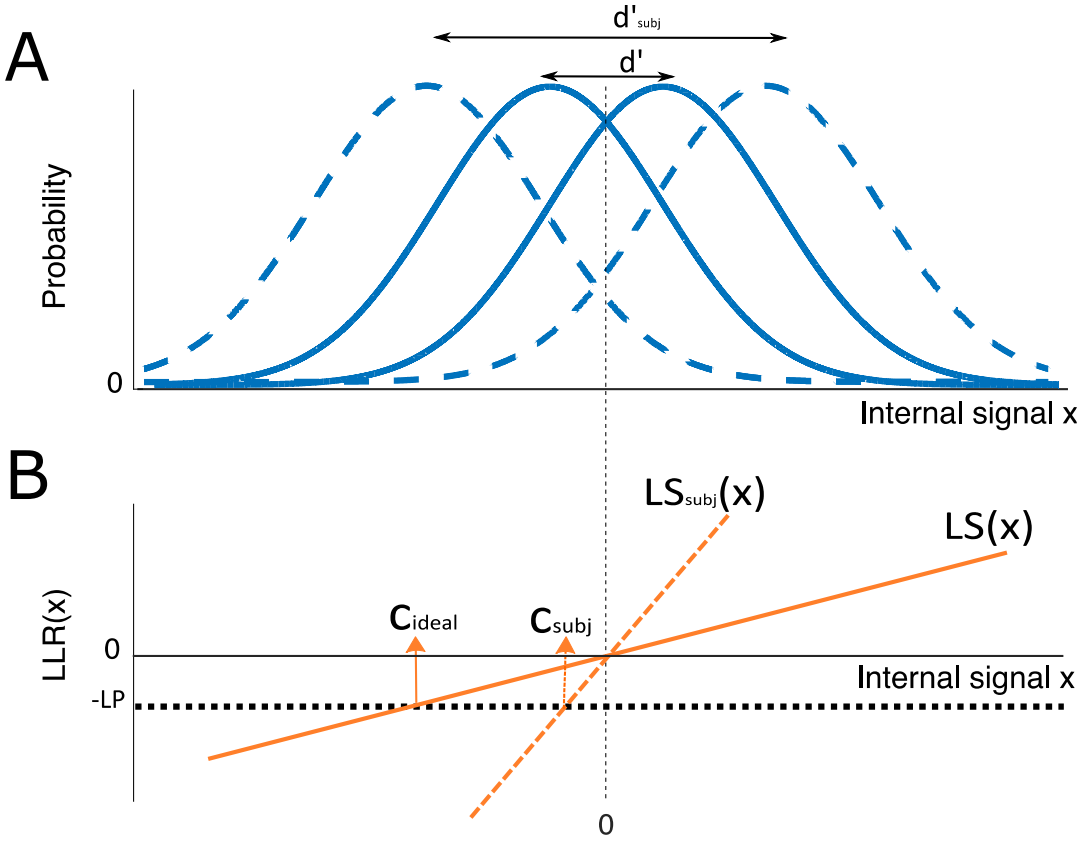
$$DV_{subj}(c_{subj}) = 0 \Leftrightarrow LP + \ d'_{subj}c_{subj} = 0 \Leftrightarrow c_{subj} = -\frac{1}{d'_{subj}}LP \qquad \text{(eq. 4b)}$$

Rearranging equations 4a and 4b, we obtain the fundamental prediction of our theoretical model defined as:

$$c_{subj} = c_{ideal}\frac{d'}{d'_{subj}} \quad \text{(eq. 5)}$$

Equation 5 states that the criterion set by the non- ideal observer should deviate from the ideal criterion by a factor that is the inverse of her overconfidence bias, leading to a conservative decision bias if she is overconfident (i.e., $\frac{c_{subj}}{c_{ideal}} < 1 \; if \; d'_{subj} > d'$) and a liberal decision bias if she is underconfident (i.e., $\frac{c_{subj}}{c_{ideal}} > 1 \; if \; d'_{subj} < d'$).

Figure 2. *Signal Detection Theory model*



Note. (A) Probability distributions of internal signal x for the ideal observer and the non-ideal observer. The ideal observer (blue curves) perfectly estimates her internal signal, whereas the non-ideal observer (blue dotted curves) overestimates the quality of her internal signal resulting in overconfidence bias (i.e., $d'_{subj} > d'$). (B) Log Likelihood Ratio for a given internal signal x for the ideal observer (full orange line) and the non-ideal observer (dotted orange line). The ideal observer sets her decision criterion when $LS(x) = -LP$, whereas the non-ideal observer sets her decision criterion when $LS_{subj}(x) = -LP$ resulting in conservative decision bias (i.e., $\frac{c_{subj}}{c_{ideal}} < 1$) in the case of overconfidence.

2.3.2. Empirical strategy

To test this model empirically, we first estimate the right-hand side of equation 5. As detailed below, c_{ideal} is estimated during the cueing session, and the ratio $\frac{d'}{d'_{subj}}$ is estimated from the confidence session thus the right-hand side of this equation can be fully determined. Then, we compare the predicted criterion (c_{subj}) to the actual criterion used by participants in the cueing session (c_{obs}).

2.3.2.1. *Estimating overconfidence bias from the confidence session*

In the confidence session, the prior probabilities of the two states were equal therefore $LP = 0$. The link between subjective and objective probabilities follows from equations 3a and 3b:

$$\log\left(\frac{P(A|x)}{P(B|x)}\right) = \frac{d'}{d'_{subj}} \log\left(\frac{P_{subj}(A|x)}{P_{subj}(B|x)}\right) \quad \text{(eq. 6)}$$

We grouped the trials into subsets according to the confidence reported (i.e., 50, 60, 70, 80, 90, or 100%) and response (i.e., left, or right), and, for each subset, we evaluated the subjective probability P_{subj} (i.e., the average confidence) and the objective probability P (i.e., the actual frequency) of a given state. Converted in log-odds, these quantities provide an estimation of $\log\left(\frac{P_{subj}(A|x)}{P_{subj}(B|x)}\right)$ and $\log\left(\frac{P(A|x)}{P(B|x)}\right)$, respectively. According to equation 6, overconfidence bias (as defined in the model) can thus be estimated by the inverse of the coefficient of the linear regression of subjective probabilities over objective probabilities (both expressed in log-odds) (see Fig S1 in the Supplementary Materials).

2.3.2.2. *Predicting the decision criterion in the cueing session*

In the cueing session, the prior probabilities of the two states varied on a trial-by-trial basis. For each participant, we computed the observed (c_{obs}), ideal (c_{ideal}) and predicted (c_{subj}) criterion adjustment in response to unequal base rates. To evaluate the observed criterion adjustment, we fitted with maximum likelihood a SDT model to the data from the cueing session, separately for each participant. This fitted model had 4 parameters: a constant sensitivity d' and a decision criterion for

each of the 3 types of cues (left, right, neutral). To estimate the SDT parameters, we chose arbitrarily to define the state of nature A (i.e., right circle has more dots) as the Signal and the state of nature B (i.e., left circle has more dots) as the Noise. We expected the criterion for the left cue trials to be positive (i.e., corresponding to answering "right" less often than "left") and the criterion for the right cue trials to be negative (i.e., corresponding to answering "right" more often than "left"). We used the semi-distance between the estimated criteria for the left cue and right cue trials as a measure of the actual criterion adjustment ($c_{obs} = (c_{obs,left} - c_{obs,right})/2$) for each participant. The ideal criterion adjustment (for the left cue trials) is given by the relation $LP + d'c_{ideal} = 0$, with $LP = \log(.25/.75)$ in the case of a 75% valid cue. The predicted criterion adjustment (for the left cue trials) was derived using equation 5. Note that we assumed here that overconfidence bias (i.e., the ratio d'_{subj}/d' estimated in the confidence session) was identical between the confidence session and the cueing session. Finally, we also computed the value of the ideal and observed decision criterion in log odds with $c_{ideal,LO} = c_{ideal}\ d' = -LP$ and $c_{obs,LO} = c_{obs}\ d'$

2.3.2.3 *Reliability*

We evaluated the reliability of all the measures that we correlated at the individual level (see Table S1 in the Supplementary Materials), in terms of internal consistency for measures administered in one session and test-retest reliability for measures repeated in the two sessions. We used permutation-based split-half Spearman-Brown coefficients and intraclass correlation coefficients to estimate internal consistency and test-retest reliability respectively, as recommended by Parsons et al. (2019). Internal consistency of the two measures that we compared to test our model (i.e., the actual criterion and the predicted criterion) was quite good (median = 0.9305, 95%HDI= [0.8066, 0.9524] and median=0.8910; 95% HDI= [0.8186,0.9361], respectively).

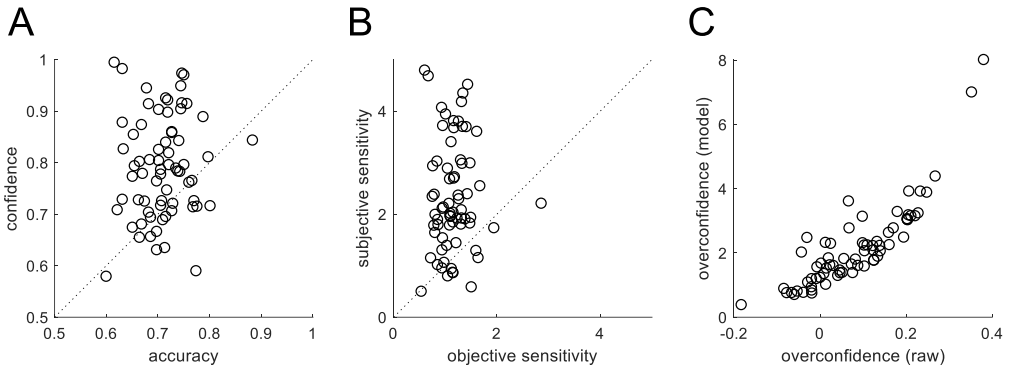
3. Results

To anticipate our results, we first established both overconfidence bias and conservative decision bias in our participants. Then, we evaluated the hypothesized link between these two measures, but critically found no evidence for this link. We thus explored whether overconfidence bias would affect perceptual performance in the cueing session in other ways and found that overconfident participants exhibited a lower sensitivity following a 75% valid cue. Model comparison then clearly confirmed that this model was much more probable given our data. An intuition for the mechanism underlying this result is provided in the discussion.

3.1. *Model-based measure of overconfidence bias*

We first evaluated participants' confidence in their perceptual decisions. Raw overconfidence bias, computed from the confidence session as the average confidence minus accuracy (Fig 3A) was largely heterogeneous across participants (M=0.08, SD=0.11), but highly significant at the group level (T-test vs. 0: t(68)=6.42, p<0.001). We also calculated overconfidence bias with our model-based measure (see Methods and Fig S1 in the Supplementary Materials), in which we used confidence ratings to quantify participants' subjective estimate of their own sensitivity (d'_{subj}), which can be compared to the actual sensitivity (d'). We found that subjective estimations of sensitivity (Fig 3B) were twice as large as actual sensitivities (ratio d'_{subj}/d': M=2.18, SD=1.31). This model-based measure of overconfidence bias was significant at the group level (T-test vs. 1: t(68)= 7.53, p<0.001), and highly correlated with the initial raw overconfidence bias across participants (r=0.85, p<0.001) (Fig 3C).

Figure 3. *Overconfidence bias in the confidence session*



Note. (A) Average confidence and average accuracy for each participant. (B) Subjective and objective sensitivity for each participant. (C) The relation between our model-based measure of overconfidence bias (i.e., ratio subjective sensitivity over objective sensitivity) and raw overconfidence bias (i.e., average confidence minus average accuracy). Each dot is a participant (N=69). In panels A and B, the black dotted line corresponds to the 45-degree line.

3.2. *Model-based measure of conservative decision bias*

We then turned to the cueing session to quantify conservative decision bias by assessing how observers combined the symbolic cue information with their sensory information. Descriptive statistics are reported in Table 1.

Table 1. *Descriptive statistics in the cueing session*

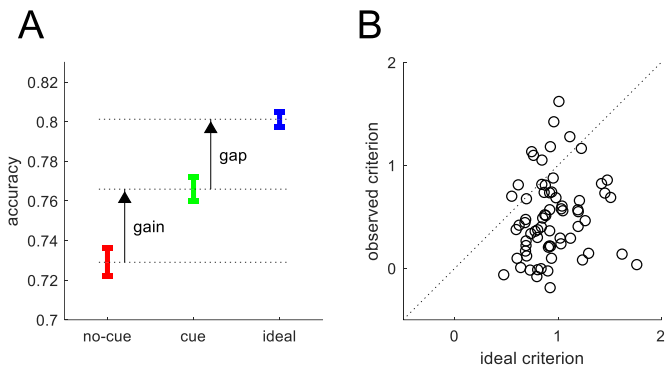
	Response rate right	Accuracy All trials	Accuracy valid cue trials	Accuracy invalid cue trials	SDT parameters with MLE		SDT parameters	
					c	d'	c	d'
Neutral cue	0.508 (0.066)	0.729 (0.058)	-	-	-0.029 (0.203)		-0.029 (0.204)	1.264 (0.364)
Left predictive cue	0.253 (0.118)	0.764 (0.057)	0.840 (0.103)	0.535 (0.207)	0.494 (0.481)	1.242 (0.344)	0.497 (0.496)	1.213 (0.460)
Right predictive cue	0.743 (0.108)	0.768 (0.060)	0.841 (0.098)	0.549 (0.190)	-0.477 (0.427)		-0.479 (0.440)	1.213 (0.426)

Note. Shown are, per cue (neutral, left and right): Response rate right, Average accuracy rate, Average accuracy rate conditional on valid and invalid cues, Equal variance SDT parameters fitted with Maximum Likelihood (the fitted model had 4 parameters: a constant sensitivity, and a decision criterion that was free to vary between the 3 conditions (neutral cue, left predictive cue, and right predictive cue), Equal variance SDT parameters estimated for each cue condition and averaged across participants' point estimates, with decision criterion c=-0.5*[Z(H)+Z(F)] and sensitivity d'=Z(H)-Z(F) where Z(H) and Z(F) are the inverse of the cumulative Gaussian distribution function for the Hit (H) and False-alarm (F) rates. Standard deviations are reported between parentheses.

Overall, participants benefited from the cue information: their performance was higher after a predictive cue compared to a neutral cue (predictive: M=0.77, SD=0.05; neutral: M=0.73, SD=0.06; accuracy gain: M=0.04, SD=0.04, t(68)=7.12, p<0.001; Fig 4A). Nonetheless, when compared to an ideal observer optimally integrating the cue while maintaining sensitivity constant (for which accuracy would be given by: $P(B)\Phi(c_{ideal} + d'/2) + P(A)\big(1 - \Phi(c_{ideal} - d'/2)\big)$ where Φ is the cumulative of the standard normal distribution), participants were not fully benefitting from the cue, as revealed by a significant accuracy gap (ideal accuracy: M=0.80, SD=0.03, accuracy gap: M=0.04; SD=0.03, t(68)=9.07, p<0.001; Fig 4A). To evaluate how participants used the cue to adjust their responses, we also compared how participants placed their decision criterion, relative to the ideal placement (Fig 4B). Ideally, participants should have adjusted their decision criteria to incorporate the information provided by the cue (in log-odds: $c_{ideal,LO}$= log(0.75/0.25)≈1.1). However, they only adjusted their criteria half-way through this ideal value ($c_{obs,LO}$: M=0.58,

SD=0.48), resulting in a significant under-adjustment (i.e., conservative decision bias) (ratio observed criteria over ideal criteria: M=0.53, SD=0.44, T-test vs. 1: t(68)=-8.99, p<0.001).

Figure 4. *Performance in the cueing session*



Note. (A) Average performance across participants, in the presence of a neutral cue (no-cue condition), a 75% valid cue (cue condition), and for ideal observers perfectly integrating the 75% valid cue (ideal condition). Error bars represent SEM. (B) Observed criterion (c_{obs}) and ideal criterion (c_{ideal}) for each participant (N=69). The black dotted line corresponds to the 45-degree line.

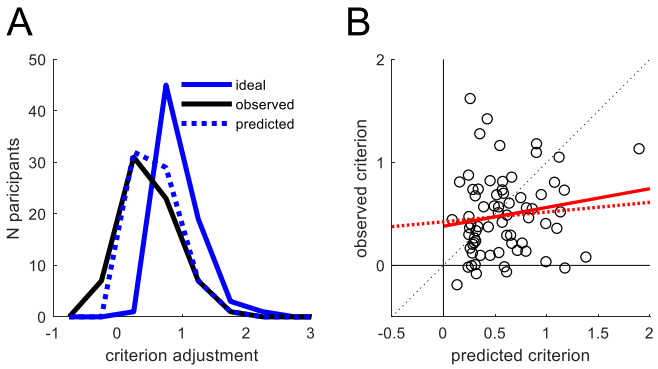
3.3. *Discarding the conservative decision bias mechanism*

We then examined whether overconfidence bias and conservative decision bias would relate, as expected under the SDT model (see Methods). At the group level, the overall amount of conservative decision bias appeared in line with the overall amount of overconfidence bias in our data: the criterion adjustment observed (c_{obs}) (M= 0.49, SD= 0.39) largely overlapped with the criterion adjustment (c_{subj}) (M= 0.58, SD= 0.34) that was predicted from participants' overconfidence bias (Fig 5A), and we could not reject at the 5% significance level the hypotheses that c_{subj} and c_{obs} have the same median (Wilcoxon rank sum test: p=0.106) or the same dispersion (Ansary- Bradley test: p=0.07).

To our surprise, however, the prediction of the SDT model did not hold when examining the covariations of predicted and actual criteria across participants. Using a one-sided Pearson

correlation analysis based on the alternative hypothesis that the predicted criterion c_{subj} has a positive correlation with the observed criterion c_{obs}, we could not reject the null hypothesis at a 5% significance level (r=0.16, p=0.094; and r= 0.0705, p-value=0.272 after we removed one outlier identified when plotting the data, see Fig 5B). With Bayesian testing, using the statistical software JASP (JASP Team, 2022), this was accompanied by a Bayes factor BF_{0+} suggesting that there is moderate evidence supporting the lack of a relationship between these two quantities. Specifically, assuming that any positive Pearson correlation coefficient ρ was equally likely a priori (i.e., using a Stretched beta prior width κ=1, truncated to allow only values between 0 and 1), BF_{0+} indicated that the data were 3.807 more likely under the null H_0 (i.e., $\rho = 0$) than the alternative directional hypothesis H_+ (i.e., $\rho \sim U[0,1]$) (see Fig S2A in the Supplementary Materials). Furthermore, we performed a robustness check to assess the sensitivity of our findings to a wide range of priors (see Fig S2B in the Supplementary Materials). We found that, the Bayes factors BF_{0+} consistently indicated moderate evidence for H_0 over H_+ for prior widths κ greater than or equal to 0.66, and indicated only anecdotal evidence for H_0 for lower prior widths (i.e., corresponding to assigning more mass to small correlation coefficients). This analysis suggests that we can discard at least a medium-to-large correlation, and that a larger sample might be needed to evaluate the possible presence or absence of a small correlation.

Figure 5. *Overconfidence bias and conservative decision bias*



Note. (A) Distribution of the adjustment of decision criteria in the presence of a symbolic cue, as observed empirically (c_{obs}, black line), predicted theoretically for ideal observer (c_{ideal}, blue line) and for overconfident participants (c_{subj}, blue dotted line). (B) The relation between the criteria observed empirically (c_{obs}) and the criteria predicted for overconfident participants (c_{subj}). Each dot is a participant (N=69). The red line represents the best-fitting regression when all observations are included. The red dotted line represents the best fitting regression after removing one outlying data point (located at x=1.9). The black dotted diagonal line corresponds to the predicted relation between the two variables.

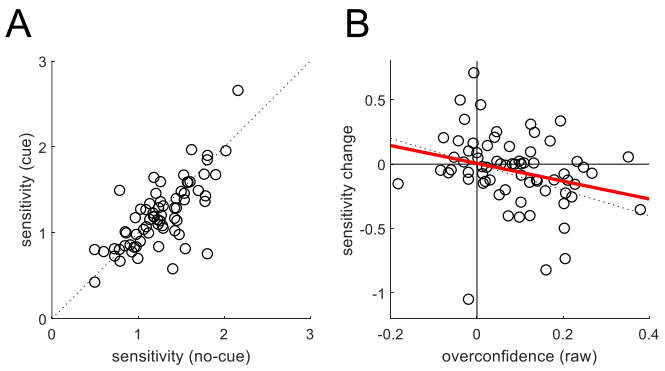
3.4. *Exploratory analysis of the sensitivity loss mechanism*

Since overconfidence bias did not lead to conservative decision bias, we conducted an exploratory analysis to evaluate whether overconfidence bias might affect performance in the cueing session via sensitivity instead. We asked whether overconfidence bias could induce a reduction in perceptual sensitivity in the 75% valid cue condition relative to the neutral condition, referred to as "sensitivity loss". This would be consistent with the idea that overconfident participants invested less effort into processing the stimuli when offered a predictive cue, perhaps because they would believe that they were already doing well enough.

To evaluate this possibility, we estimated separate values of sensitivity (d') and decision criterion (c) for each cue condition (right, left, and neutral) (see Table 1). Although there was no systematic change in sensitivity in the 75% valid cue condition relative to the neutral cue (average of d'_{right} and d'_{left}: M=1.21, SD=0.39; neutral: M=1.26, SD=0.36; difference: M=-0.05, SD=0.28, t(68)=-

1.51, p=0.13, Fig 6A), we observed a large heterogeneity across participants, which was a potential leverage to understand the relation between overconfidence bias and performance. We thus evaluated how this sensitivity change was correlated with raw overconfidence bias and found a significant negative correlation between the two measures (r=-0.265, p=0.028). In other words, more overconfident participants tended to exhibit lower sensitivity when given a predictive cue than in the no-cue condition (Fig 6B), although we should point out that overconfidence bias only explained about 7% of the variance in sensitivity change. We note that if overconfident participants have a lower sensitivity in the 75% valid cue condition, the acceleration of responses (measured as the difference in median response times between neutral and predictive cues) was not correlated with overconfidence bias (r=-0.07, p=.57). Therefore, sensitivity loss in overconfident participants was not the result of a speed-accuracy trade-off.

Figure 6. *Overconfidence bias and sensitivity loss*



Note. (A) Sensitivity in the presence of a 75% valid cue and in the no-cue (i.e., neutral cue) conditions. (B) The relation between sensitivity change (sensitivity 75% valid cue minus sensitivity neutral cue) and raw overconfidence bias. The red line represents the best-fitting regression. In both panels, each dot is a participant (N=69). The black dotted line corresponds to the 45-degree line.

To ensure that this result was not due to individual differences in global motivation to engage in the experiment, we evaluated the relation between overconfidence bias and sensitivity

change in a regression with control variables. To control for motivation to do well in the perceptual task, we used the calibrated difference in the number of dots (i.e., the difference in the number of dots between the left and right circles to calibrate stimulus difficulty per participant) averaged across both sessions, which we believe would be greater for less motivated participants. To control for motivation to do well in the confidence task, we included the resolution of confidence as well as the median of response times of confidence ratings, which we believe would be lower for less motivated participants. In addition, we controlled for cognitive abilities by taking the average value of the two working memory scores measured at the beginning of each session (see Methods S1 in the Supplementary Materials). Adding such control variables did not change our results (see Table S2 in the Supplementary Materials), suggesting that global motivation is not a confounding factor. However, we noted that the internal consistency of sensitivity change was poor (median= 0.0612, 95% HDI= [-0.3036, 0.3478]), due to the low reliability of d'_{right} and d'_{left} (median= 0.5188, 95% HDI= [0.3285, 0.6622] and median=0.5627, 95%HDI= [0.3973, 0.6940], respectively) whereas the reliability of $d'_{neutral}$ was acceptable (median =0.7771, 95% HDI= [0.6924, 0.8449]). Such low values might be attributed to the low number of trials used with the split-half method to compute the sensitivity measures d'_{right} and d'_{left}.

Given that the change in sensitivity across conditions appears to be related with overconfidence, we checked whether relaxing the assumption of constant sensitivity across conditions would affect our empirical test of the relationship between overconfidence bias and conservative decision bias. More specifically, we used the values of sensitivity and criterion estimated separately for each cue condition to compute again the observed criterion (M= 0.49, SD= 0.40) and the predicted criterion (M= 0.65, SD= 0.49). Still, we found no evidence in support of a positive correlation between these two measures at a 5% significance level (r= 0.17, p-value=0.085 after we removed two outliers identified when plotting the data) and found that the plots of the data look very similar (for comparison see Fig S3 in the Supplementary Materials), suggesting that the variation in sensitivity observed in our data did not affect our findings.

3.5. *Model comparison: conservative decision bias vs. sensitivity loss mechanisms*

We used a model comparison approach to evaluate which mechanism better describes participants' behavior: conservative decision bias as hypothesized initially or sensitivity loss as suggested by our exploratory analysis (see Table 2). In a series of probit mixed-effects models (DeCarlo, 1998; Knoblauch & Maloney, 2012), we estimated how participants' responses (in the cueing session) were predicted on a trial-by-trial basis by the stimulus and the cue presented in each trial, and their interaction with participants' raw overconfidence bias (calculated in the confidence session). Our simplest model (Model 0) included no effect of overconfidence bias but only effects of stimulus and predictive cue. Note that, in Table 2, the coefficients we obtain for Model 0 correspond to the SDT estimates. In particular, the *Intercept* is an estimate of the SDT-criterion in the presence of a neutral cue, the coefficient of *CuePred* is an estimate of the shift in criterion in presence of a predictive cue (as opposed to a neutral cue), and the coefficient of *Stimulus* is an estimate of the sensitivity. In addition, Model 1 included a two-way interaction *CuePred x Overconf* to allow overconfidence bias to affect the criterion placement in presence of a predictive cue, thereby implementing the conservative decision bias mechanism predicted initially, while Model 2 included a three-way interaction *CuePred x Stimulus x Overconf* to allow overconfidence bias to affect the sensitivity in presence of a predictive cue, thereby implementing the sensitivity loss mechanism identified in our explanatory analysis.

In Model 1, the two-way interaction term *CuePred x Overconf* is not significantly different from 0 thus overconfidence bias does not produce a significant difference in criterion placement in presence of a predictive cue. On the other hand, in model 2, the three-way interaction term is significantly different from 0 at a 5% significance level and the sign is negative, bringing support to the novel hypothesis that overconfidence bias reduces sensitivity in presence of a predictive cue.

Comparing nested models using likelihood ratio tests, we found that including the modulation of sensitivity by overconfidence better described the data (model 2 vs. model 0: $\chi^2(5) =$

16.886, p<0.01; and model 2 vs. model 1: $\chi^2(3) = 12.551$, p<0.01). However, including the modulation of the criterion by overconfidence did not (model 1 vs. model 0: $\chi^2(2) = 4.335$, p = 0.115). Comparing the Akaike Information Criteria (AIC), provides evidence, as well, in favor of model 2 (i.e., the "sensitivity loss" mechanism). According to the raw AIC values, model 2 is the preferred model since it has the lowest AIC value (Model 0: AIC= 37166.9, Model 1: AIC= 37166.57, Model 2: AIC=37160.02). In addition, comparing the Akaike weights, we found that model 2 is $\frac{w_2(AIC)}{w_1(AIC)}$=26.7 times more likely to be the best model in a Kulback-Leibler sense than is the next best fitting model (model 1).

Table 2*. Model comparison*

	DV: Response		
	Model 0	Model 1	Model 2
Intercept	0.029 (0.024)	-0.009 (0.030)	-0.007 (0.030)
CuePred	0.498*** (0.046)	0.464*** (0.058)	0.465*** (0.058)
Stimulus	1.221*** (0.041)	1.221*** (0.041)	1.259*** (0.054)
Overconf		-0.456* (0.217)	-0.434* (0.220)
CuePred x Overconf		0.399 (0.426)	0.493 (0.424
CuePred x Stimulus			-0.027 (0.040)
Stimulus x Overconf			-0.038 (0.390)
CuePred x Stimulus x Overconf			-0.671* (0.294)
Random effects	YES	YES	YES
Df_i	9	11	14
LL_i	-18574.5	-18572.3	-18566
AIC_i	37166.9	37166.57	37160.02
w_i(AIC)	0.030	0.035	0.935
Nb observations	35328	35328	35328
Nb participants	69	69	69

Note. Each model is a probit regression in which participants' responses were predicted on a trial-by-trial basis. To account for random variations across participants, all models included random criterion effects and random sensitivity effects at the participant's level. *Response* is coded as 1 if the participant responds "left" and 0 if he/she responds "right". *Intercept* is coded as -1. *Stimulus* is

coded as -0.5 if the stimulus category is right and +0.5 if the stimulus category is left. *CuePred* is coded as 1 for predictive cues and 0 for neutral cues. Note that to study the effect of CuePred on participants' responses, we grouped the trials in which a right or left cue was presented and reversed the coding of the variables *Response* and *Stimulus* when a right cue was presented. *Overconf* is the participant's raw overconfidence bias measured in the confidence session (i.e., average confidence minus average accuracy). Df_i is the degree of freedom for model *i*. LL_i is the logarithm of the maximum likelihood for model *i*. AIC_i is the Akaike information criterion for model *i.* $w_i(AIC)$ is the rounded Akaike weight for model *i*. Estimations are performed with the glmer function of the R package lme4 (Bates et al., 2015). Standard errors are reported in parentheses. pvalues: *p<0.05; **p<0.01; ***p<0.001.

4. Discussion

We described a theoretical model according to which overconfidence bias would induce conservative decision bias in response to unequal base rates. Within the SDT framework, we provided a quantitative measure of overconfidence bias (i.e., the overestimation of one's own sensitivity to the sensory signal) and derived a predicted criterion adjustment from this value. We then tested the proposed model using a psychophysical task. Contrary to what was prescribed by the model, overconfidence bias and conservative decision bias appeared uncorrelated across participants. Our data prompted us to consider that overconfidence bias may affect performance via a different mechanism. And, our final analysis suggested that overconfidence bias may induce a reduction in sensitivity in response to unequal base rates. Specifically, we found that overconfidence bias was positively correlated with this sensitivity loss. Model comparison confirmed this novel finding, that participants' decisions are better explained by an effect of overconfidence bias on sensitivity rather than on criterion adjustment.

That overconfidence bias and conservative decision bias were uncorrelated is consistent with a recent study (Ackermann & Landy, 2015) that found that misestimating internal response variability (formally equivalent to our d'_{subj}) in a visual detection task is not the cause of conservative decision bias. Our findings are similar but by collecting confidence data we provide a more direct test of the hypothesized link. It should be emphasized that this absence of correlation is not likely to be due to poor experimental measures of the two quantities, as both have good internal

consistency and were clearly manifest in our data, in line with previous studies using perceptual tasks reporting overconfidence bias (Baranski & Petrusic, 1994; Kubovy, 1977; Kvidera & Koutstaal, 2008; Mamassian, 2008; Massoni et al., 2014) and conservative decision bias (Ackermann & Landy, 2015; Gorea & Sagi, 2000; Green & Swets, 1966; Kubovy 1977; Morales et al., 2015). Furthermore, although we cannot verify in our data that overconfidence bias remained stable across the two experimental sessions spaced four days apart, as we hypothesized in our empirical strategy, recent studies (Ais et al., 2016; Navajas et al., 2017) have found evidence to support this assumption. However, we should also note that the present study relied on several assumptions that have been questioned in the literature. In particular, our estimation of conservative decision bias was based on Signal Detection Theory with the assumption of Gaussian noise (see Fig 2), and conservative decision bias observed in our data could be the result of this assumption if it were incorrect (Maloney & Thomas, 1991). In addition, we assumed that participants reported their true confidence level but, even though their confidence report was incentivized, this might not have been the case as some authors have pointed out that the mapping of internal evidence into stated probabilities can suffer from biases (e.g., Fox and Clemen, 2005; Higham et al., 2015). Similarly, even though participants were instructed to use the probabilistic information provided by the cues to maximize their earnings, we cannot discard the possibility that the mapping of the 75% probability into internal evidence might have been distorted (e.g., Zhang & Maloney, 2012).

We will now offer some tentative explanation for this correlation between overconfidence bias and sensitivity loss in response to unequal base rates. We acknowledge that this explanation is offered a posteriori and needs to be evaluated against new data in future work. Our explanation relies on two assumptions. Our first assumption states that participants face a tradeoff between the effort they deploy when they perform the task, which allows them to maintain their sensitivity, and the satisfaction they obtain by producing correct answers. When considering this tradeoff, they might evaluate that there is a level of effort that maximizes the difference between expected benefits and costs, and aim for this level. This idea is similar to ideas of rational inattention in

behavioral economics (Sims, 2003) and expected value of control in cognitive neuroscience (Shenhav et al., 2013). More generally, this idea of self-regulation is also put forward in the domain of education, to understand how students allocate their resources when preparing for an exam (Son & Metcalfe, 2000). In this scheme, participants may see the information provided about the a priori probability of occurrence of the stimuli as an opportunity to maintain performance while deploying less effort during the task, such that they would target a lower effort (resulting in lower sensitivity). Indeed, an experiment showed that participants' subjective evaluations of effort decreases when diagnostic cues are available to help them (Botzer et al., 2013). Our second assumption simply states that overconfidence bias corresponds to participants overestimating their probability of being correct when performing the task. Although this assumption is uncontroversial, it has some non-trivial effects when combined with the first assumption. Specifically, since overconfident and well-calibrated participants would not evaluate their accuracy at the same level, they will face different trade-offs, with distinct optimal solutions in terms of effort allocation. Again, we insist that, since this explanation is offered after the fact, other mechanisms could be formulated to explain the loss of sensitivity observed in our data. More theoretical work and new empirical data would be needed to uncover the mechanism by which overconfidence bias and sensitivity loss are related. Our findings thus bring new perspectives on the role of overconfidence bias on the strategic allocation of resources in such situation.

In sum, the Signal Detection Theory approach allowed us to break down participants' suboptimal decisions in response to unequal base rate into two components, namely a sensitivity loss and an under-adjustment of criterion (i.e., conservative decision bias). And, our data suggest that overconfidence bias leads to suboptimal decisions via a sensitivity loss mechanism, independently of the under- adjustment of criterion, which is also present but unrelated to overconfidence in our data. Given that overconfidence bias and conservative decision bias have been observed, although separately, for a diverse range of participants with laboratory tasks using basic visual decisions but also in experiments emulating real-world decisions, we expect our finding

(i.e., the absence of a positive link between these two biases) to generalize to visual stimuli in which participants make similar discrimination tasks. A direct replication would need to calibrate the difficulty of the task, measure confidence in decision when stimuli are a priori equally likely to occur, fully inform participants about the manipulation of the base rate and incentivize them to be accurate. We have no reason to believe that the results depend on other characteristics of the participants, materials, or context. On the other hand, and to the best of our knowledge, we lack prior direct evidence supporting our finding regarding the link between overconfidence bias and change in sensitivity and given the poor internal consistency of the measure of sensitivity change the correlation that we found might differ in future replications. Finally, it must be noted that our sample size was calculated to detect a correlation between overconfidence and conservative decision bias with at least a medium effect size. If instead one assumes that this correlation exists but might be very small, then a larger sample size would be needed to demonstrate it. In addition, such a small correlation (if it existed) would diminish the practical and theoretical importance of the mechanism linking overconfidence to conservative decision bias. Thus, in any case, further investigations are needed to examine other possible sources of conservative decision bias.

References

Ackermann, J. F., & Landy, M. S. (2015). Suboptimal decision criteria are predicted by subjectively weighted probabilities and rewards. *Attention, Perception, & Psychophysics*, *77*(2), 638-658.

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition*, *146*, 377-386.

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, *55*(4), 412-428.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. Science, *329*(5995), 1081-1085.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, *9*(3), 226-232.

Botzer, A., Meyer, J., Bak, P., & Parmet, Y. (2010). User settings of cue thresholds for binary categorization decisions. *Journal of Experimental Psychology: Applied*, *16*(1), 1.

Botzer, A., Meyer, J., & Parmet, Y. (2013). Mental effort in binary categorization aided by binary cues. *Journal of Experimental Psychology: Applied*, *19*(1), 39.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433-436.

Broadway, J. M., & Engle, R. W. (2010). Validating running memory span: Measurement of working memory capacity and links with fluid intelligence. *Behavior Research Methods*, *42*(2), 563-570.

Chi, C. F., & Drury, C. G. (1998). Do people choose an optimal response criterion in an inspection task?. *IIE transactions*, *30*(3), 257-266.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic bulletin & review*, *12*(5), 769-786.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological methods*, *3*(2), 186.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175-191.

Fox, C. R., & Clemen, R. T. (2005). Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science*, *51*(9), 1417-1432.

Gorea, A., & Sagi, D. (2000). Failure to handle more than one internal representation in visual detection tasks. *Proceedings of the National Academy of Sciences*, *97*(22), 12380-12384.

Green, D. M. and Swets, J. A. (1966). *Signal detection theory and psychophysics*. Peninsula Publ, Los Altos Hills, Calif, repr. ed edition. ISBN 978-0-93214623-6.

Higham, P. A., Zawadzka, K., & Hanczakowski, M. (2016). Internal mapping and its impact on measures of absolute and relative metacognitive accuracy. *The Oxford handbook of metamemory*, 39-61.

JASP Team (2022). JASP (Version 0.16.1) [Computer software].

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist, 73*(1), 1–2.

Knoblauch, K., & Maloney, L. T. (2012). *Modeling psychophysical data in R* (Vol. 32). Springer Science & Business Media.

Kubovy, M. (1977). A possible basis for conservatism in signal detection and probabilistic categorization tasks. *Perception & Psychophysics*, *22*(3), 277-281.

Kvidera, S., & Koutstaal, W. (2008). Confidence and decision type under matched stimulus conditions: overconfidence in perceptual but not conceptual decisions. *Journal of Behavioral Decision Making*, *21*(3), 253-281.

Levitt, H. C. C. H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical society of America*, 49(2B), 467-477.

Maloney, L. T., & Thomas, E. A. (1991). Distributional assumptions and observed conservatism in the theory of signal detectability. *Journal of Mathematical Psychology*, *35*(4), 443-470.

Mamassian, P. (2008). Overconfidence in an objective anticipatory motor task. *Psychological Science*, *19*(6), 601-606.

Massoni, S., Gajdos, T., & Vergnaud, J. C. (2014). Confidence measurement in the light of signal detection theory. *Frontiers in Psychology*, *5*, 1455.

Morales, J., Solovey, G., Maniscalco, B., Rahnev, D., de Lange, F. P., & Lau, H. (2015). Low attention impairs optimal incorporation of prior knowledge in perceptual decisions. *Attention, Perception, & Psychophysics*, *77*(6), 2021-2036.

Murrell, G. A. (1977). Combination of evidence in a probabilistic visual search and detection task. *Organizational Behavior and Human Performance*, *18*(1), 3-18.

Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P. E., & Bahrami, B. (2017). The idiosyncratic nature of confidence. *Nature human behaviour*, *1*(11), 810.

Parsons, S., Kruijt, A. W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378-395.

Pollack, I., Johnson, L. B., & Knaff, P. R. (1959). Running memory span. *Journal of experimental* Psychology, 57(3), 137-146.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software].

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, *41*.

Rahnev, D., Lau, H., & De Lange, F. P. (2011). Prior expectation modulates the interaction between sensory and prefrontal regions in the human brain. *Journal of Neuroscience*, 31(29), 10741-10748.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217-240.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 204.

The Math Works, Inc. (2014). MATLAB (Version 2014a) [Computer software].

Ulehla, Z. J. (1966). Optimality of perceptual decision criteria. *Journal of Experimental Psychology*, *71*(4), 564.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779-804.

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human Factors*, *51*(3), 281-291.

Yang, T., & Shadlen, M. N. (2007). Probabilistic reasoning by neurons. Nature, *447*(7148), 1075

Zeiliger, R. (2000). A presentation of regate, internet based software for experimental economics. Retrieved from http://www.gate.cnrs.fr/ zeiliger/regate/RegateIntro.ppt.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in neuroscience*, *6*, 1.

Electronic supplementary material

The file Supplementary Materials contains the following items:

- Methods S1. Running memory span task
- Table S1. Reliability estimates of the measures
- Table S2. Robustness check: link between overconfidence bias and sensitivity loss
- Figure S1. Individual estimation of the model-based measure of overconfidence bias
- Figure S2: Bayesian Pearson correlation analysis
- Figure S3. The relation between the criteria observed empirically and the criteria predicted for overconfident participants when relaxing the assumption of constant sensitivity