The statistical signature of confidence is not necessarily a folded X-pattern

Manuel Rausch[1] and Michael Zehetleitner[1]

Catholic University of Eichstätt-Ingolstadt

Author Note

[1] Katholische Universität Eichstätt-Ingolstadt, Fakultät für Psychologie und Pädagogik, Fachgebiet Psychologie II, Eichstätt, Germany.

The research presented here was in part funded by the Profor+ internal research funding of the Catholic University of Eichstätt-Ingolstadt (to MR) and in part by the Deutsche Forschungsgemeinschaft (grants ZE 887/8-1 to MZ and RA2988/3-1 to MR). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Correspondence should be addressed to: Manuel Rausch, Katholische Universität Eichstätt-Ingolstadt. Psychologie II, Ostenstraße 25, 85072 Eichstätt, Germany. E-Mail: manuel.rausch@ku.de.

Abstract

Confidence in perceptual choices is a degree of belief that a choice about a stimulus is correct. To identify the neural correlates of decision confidence, recent studies have widely used statistical signatures of confidence. The most widely used statistical signature is the folded X-signature, which states that the subjective probability of being correct is 0.75 when the stimulus is neutral about the choice, increases with discriminability in correct trials, and decreases with discriminability in incorrect trials. We show that the folded X-signature is limited to specific conditions. If decision makers are provided with evidence about discriminability, objective confidence follows a different statistical signature: for both correct and incorrect choices, confidence increases with discriminability. In addition, if the simulated experiment involves discrete levels of discriminability, confidence in choices about neutral stimuli is not 0.75. Overall, this means that researchers should not search for correlates of confidence by assuming the folded X-signature a priori.

Introduction

Confidence is a metacognitive evaluation of decision making: Each choice can be accompanied by some degree of confidence that the choice is correct. In neuroscience, confidence has become a flourishing research topic, uncovering the underlying neural mechanisms in humans [1–6] as well as non-human animals [7–13]. A major obstacle to the scientific study of confidence is the inherently subjective nature of the psychological construct of decision confidence. Therefore, a large amount of recent research on confidence has been inspired by a novel approach that formalizes confidence mathematically as an objective statistical quantity [14,15]. This formalization defines confidence as the belief that a choice is correct [16]. From a Bayesian perspective, beliefs are best formalised as probabilities [17,18]. Decision confidence in this formalization is the posterior probability of being correct given the evidence [16,19]. Several predictions about objective confidence have been formally derived from the model to which we subsequently refer to as the standard model of confidence [7,14,15]: First, the average objective confidence in correct choices increases as a function of the discriminability of the stimulus. Second, the average confidence in incorrect choices decreases with discriminability. Finally, when the stimulus is neutral about the choice options, confidence is exactly 0.75. The overall pattern, which we refer to here as folded X-signature [20], has been dubbed a “statistical signature of confidence” [14,21]. Given that the folded X-signature follows objectively from the posterior probability of being correct, it has been argued that when the folded X-pattern is detected in another behavioural, neural, or physiological variable, that variable should be considered a correlate of confidence [7,14,15,22]. Thus, it is frequently used to empirically identify correlates of decision confidence [7,8,10,22–25]. Nevertheless, a recent study suggested that the Bayesian calculation of the posterior probability of being correct does not necessarily

imply the folded X-signature [26]. Likewise, the folded X-signature does not necessary imply the Bayesian calculation of confidence [27–29].

Here, we show that the folded X-signature is no longer expected when confidence is informed by a trial-by-trial representation of discriminability. When objective confidence is calculated from a model of confidence which is more general in the sense that it includes a representation of discriminability, the folded X-signature occurs only as a special case when the evidence about the discriminability of a specific stimulus is not reliable. When there is accurate information about the discriminability of a stimulus, confidence tends to increase as a function of discriminability in correct and incorrect trials, which is why we refer to this pattern as the double increase signature.

The standard model of confidence

The standard model of confidence is depicted in Fig 1. According to Sanders et al. [14], when an observer is presented with a stimulus and asked to make a choice $\vartheta \in \{-1,1\}$ about the stimulus, the stimulus d is a continuous variable that differentiates between the two options of ϑ. Negative values of d mean that observers should choose $\vartheta = -1$; $d = 0$ means no objective feature of the stimulus suggests any of the two options, and positive values indicate that observers ought to choose $\vartheta = 1$. As the sign of d determines what response observers ought to give, we refer to the sign of the stimulus as identity I. The absolute value of $|d|$ is referred to as discriminability: The greater is the distance between d and 0, the easier is the choice. The accuracy of the choice A is 1 if I and ϑ are the same, and 0 otherwise. However, observers cannot perceive d directly, instead, the choice is based on noisy sensory evidence e_I (referred to as percept by Sanders et al.), which can be considered an estimate of d. The most frequent approach is to model e_I as a random sample from a Gaussian with a mean of d, while ϑ is modelled as a deterministic function of d. Finally, given that observers know the distributions from which d and e_I are sampled, the posterior probability of a correct

choice given the sensory evidence e_I and the choice ϑ can be calculated based on Bayes' theorem (see S1 Appendix).

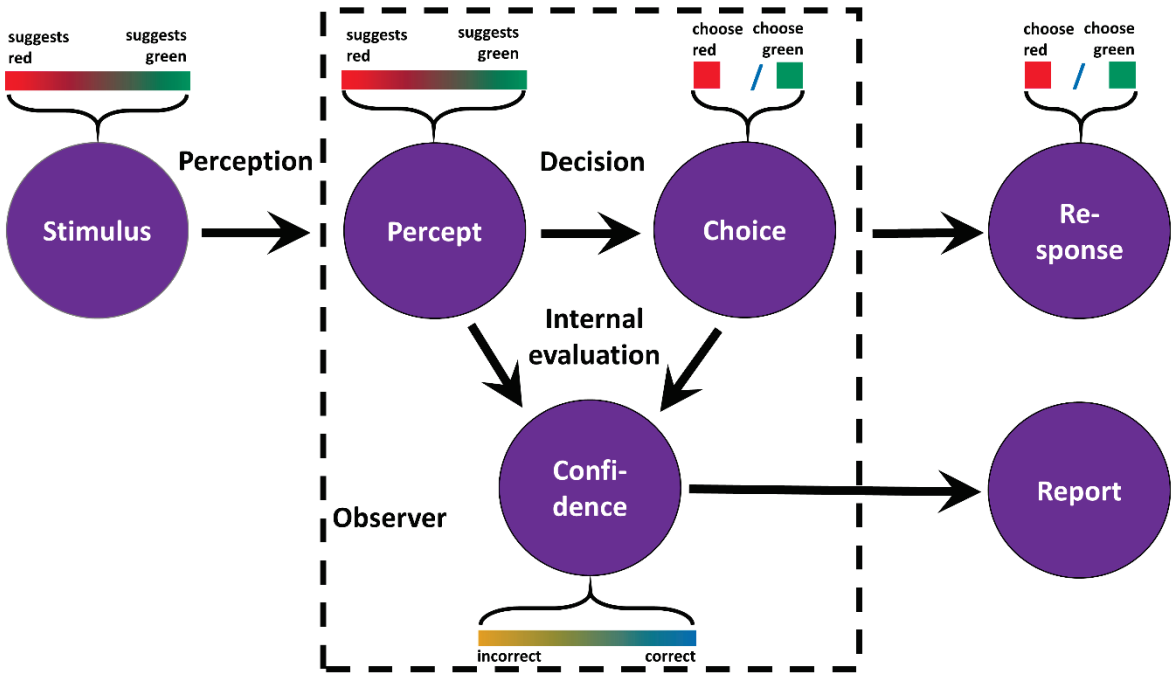

Fig 1. The standard model of confidence. The stimulus objectively supports the choice options "red" and "green" to varying degrees. As perception is noisy, the percept is a corrupted representation of the degree to which the stimulus favours a specific choice option. Confidence is the probability of making the correct choice given percept and choice.

The standard model has been presupposed to derive the folded X-signature [14,15], although different aspects of the folded X-signature come with specific additional assumptions: First, confidence in choices about neutral evidence is .75 only if the distribution of the stimulus is uniform and yields choice accuracies spanning from 0.5 to 1, if sensory evidence is sampled from a symmetric distribution with a single peak centred on the stimulus, and if choice is deterministic [14,15,26]. Second, the decrease of confidence in incorrect choices presupposes that the observer is not provided with any information about the discriminability of the stimulus at the level of single choices [15]. Although the Bayesian calculation of the probability of being correct implies knowledge of the distribution from which d is sampled, knowledge the distribution of d only implies that observers know the probability of the degrees of discriminability across the experiment. For each specific choice

however, the standard model assumes that observers do not possess any knowledge what the discriminability of the stimulus is over and above the distribution from which d is sampled.

The general model of confidence

The general model of confidence extends the standard model by including the possibility that observers perceive or infer the discriminability of the stimulus on the level of single choices. For example, when a driver in heavy rain needs to discern if a traffic light is green or red, the driver might not only be unsure because their colour percept is ambiguous, but they might also be cautious because they see or know their view is hindered by rain. Analogous to traffic lights and rain, many psychophysical experiments do not manipulate the stimulus as one independent variable; instead, two features of the stimulus are varied across the experiment. Therefore, the general model of confidence (see Fig 2) considers identity I and discriminability d as two independent aspects of each single stimulus: The identity, which in each trial can be either -1 or 1, is the variable in the external world that determines which of the choice options is correct. The model generates a choice ϑ about the identity I of the stimulus. For example, the stimulus could be red or green, and participants need to make a choice accordingly. Choices are correct when I and ϑ are both either -1 or 1. Discriminability d is the variable in the external world that determines how easy/difficult the choice is. For instance, many experiments manipulate contrast, presentation time, or luminance orthogonally to stimulus identity I. According to the general model, observers in each single trial obtain sensory evidence about *both* aspects of the stimulus, i.e. there is sensory evidence for identity e_I, and evidence for discriminability e_d. While e_I depends on I and on d, e_d depends only on d, but not on I. To represent that observers' do not have direct access to I and d, e_d is sampled from a Gaussian distribution whose mean depends on d, and e_I is sampled from a Gaussian whose mean depends on I and on d. The posterior probability

of a correct choice given and the choice ϑ can again be calculated based on Bayes' theorem (see S2 Appendix).

There are at least two possibilities why in an experimental situation, evidence about discriminability e_d may exist separately from the evidence about the identity e_I: First, when stimuli with different degrees of discriminability are not presented in random sequence, for example when discriminability is constant within one block of the experiment, observers can infer the discriminability of the present stimulus. A second possibility is that observers in many cases are able to perceive discriminability directly: Within the visual system, there is not only sensory evidence about the choice-relevant stimulus feature I, but also sensory evidence about other features of the stimulus, irrelevant to the current choice [30,31]. For example, in a masked orientation task, observers may estimate the discriminability not only by their percept of the orientation, but also by their percept of the shape, texture, or presentation time of the stimulus, even when these features are not explicitly manipulated by the experimenter [32]. All sensory evidence irrelevant to the current choice can be used as evidence about the discriminability as long as it is correlated with discriminability.

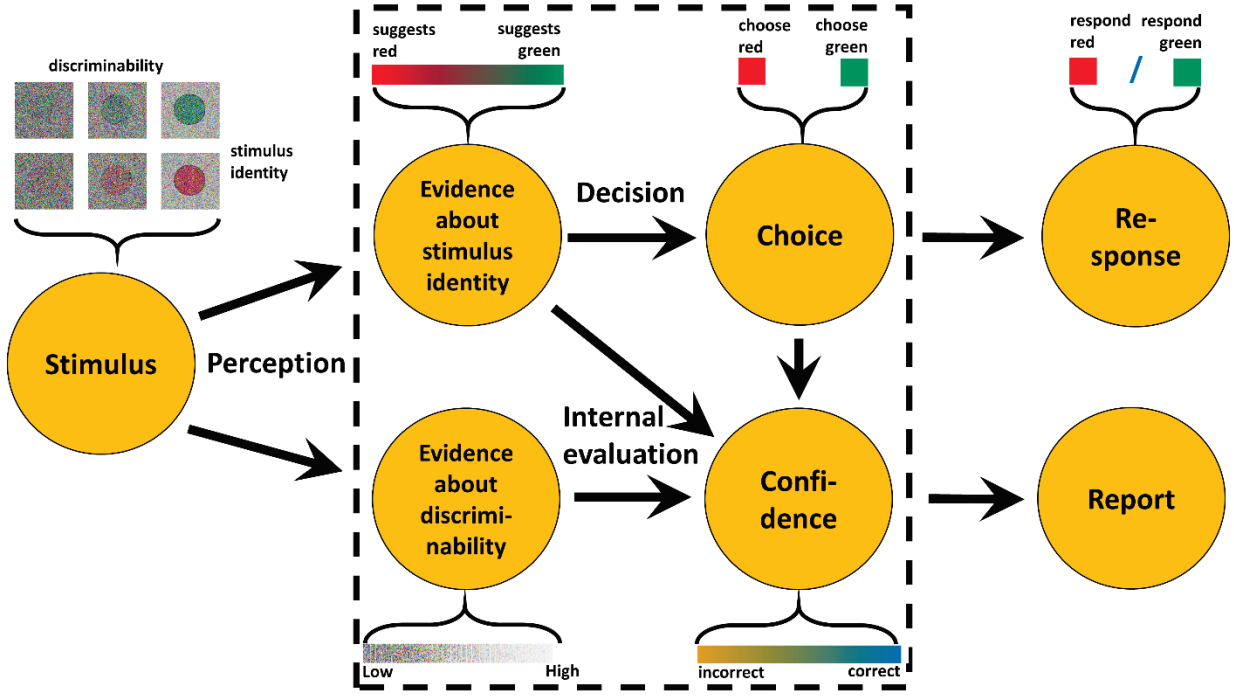

Fig 2. The general model of confidence. The general model is a generalization of the standard model. In many psychophysical experiments, the stimulus varies in two aspects: stimulus identity (symbolized here as red and green colour patches) and discriminability (symbolized here by the noise dots). In the general model, the stimulus generates two internal variables: the evidence about the stimulus identity, a continuous variable that differentiates between the possible identities, and evidence about the discriminability. Objective confidence about the correctness of the choice is based on evidence about the identity as well as evidence about discriminability.

Why is confidence not exclusively based on sensory evidence dependent on the choice-relevant features of the stimulus if decision confidence is calculated objectively, but also on evidence for the quality and reliability of perception itself? The key fact is that confidence as the posterior probability that the choice is correct given the evidence is only objective if it includes all information that is dependent on the stimulus. Given confidence is objective only if all evidence available is used, and if e_d exists in a specific task, it follows that objective confidence should be based on e_d, too.

Rationale of the present study

In the present study, we used Monte Carlo simulations to trace the statistical signatures of optimal confidence calculated as the posterior probability of being correct given the evidence. Our simulations were based on the standard model as well as on the general model, which extends the standard model by assuming that observers on single trial basis obtain evidence about the discriminability of the stimulus. Based on the general model, we also examined the impact of the reliability of evidence about discriminability on the statistical signature of confidence. Finally, we examined if relying confidence on evidence about discriminability is a beneficial strategy, or if it is an example of a suboptimal mental shortcut to the probability of being correct [6,27,33–35], i.e. a heuristic [36,37].

Results

Standard model

Fig 3 shows the statistical signatures of confidence obtained from simulations based on the standard model. Only two of the three postulated features of the folded X-signature consistently follow from the standard model: Independent of the distribution of discriminability |d|, confidence in correct choices always increases as a function of discriminability, and confidence in incorrect choices always decreases with discriminability. However, when the stimulus is neutral about the choice options, confidence is .75 only when |d| is sampled from a continuous uniform distribution that includes high discriminability (see Fig 3f). When |d| is sampled from a discrete uniform distribution (Fig 4a-c) or a gamma distribution (Fig 3d, e), or when the continuous uniform distribution does not support high discriminability (Fig 3a, b), confidence in choices about neutral stimuli is not .75.

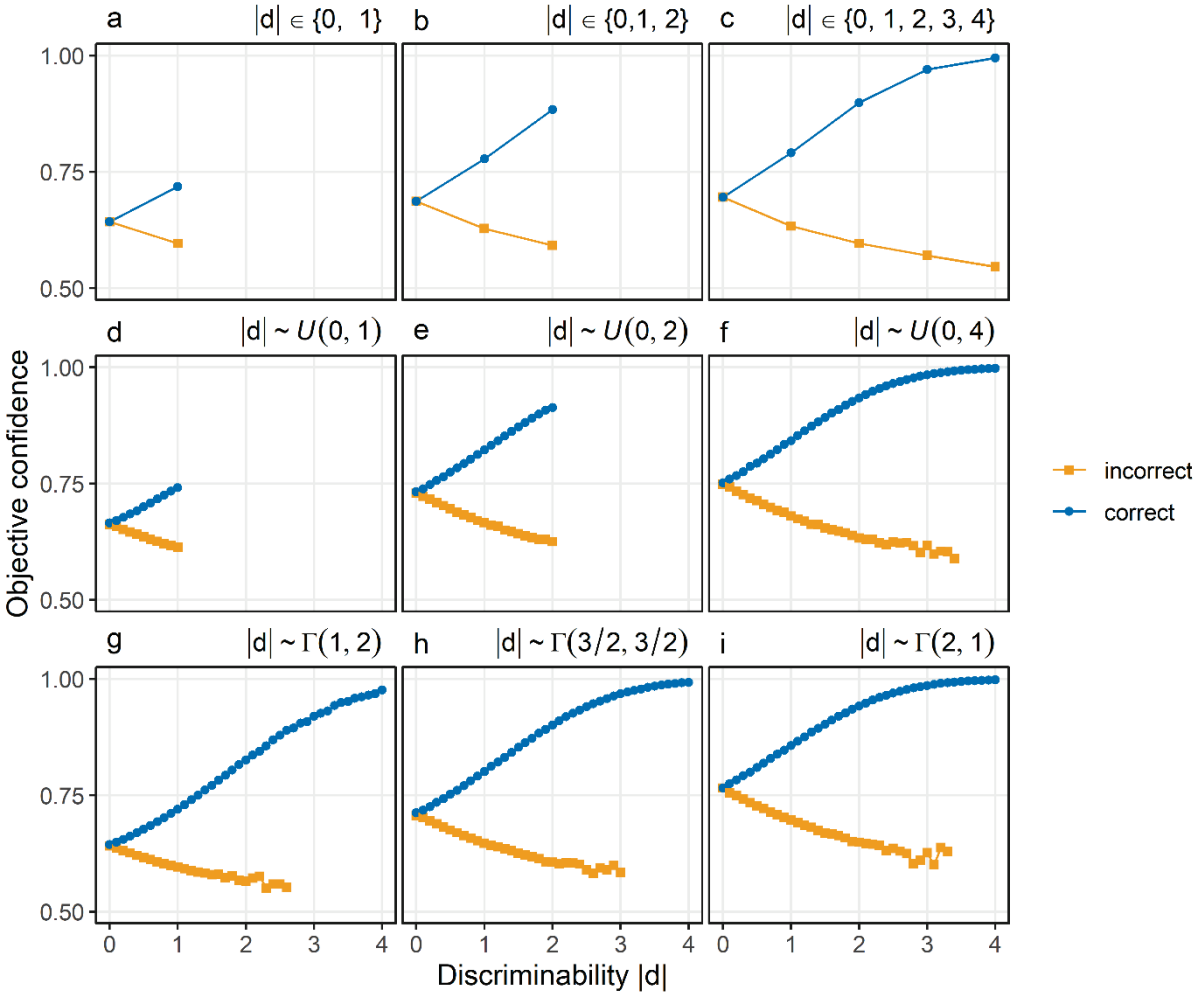

Fig 3. Objective confidence given the standard model of confidence. Confidence (y-axis) is shown as a function of discriminability (x-axis) in correct choices (blue) and incorrect choices (orange). Different panels show different distributions from which discriminability was sampled. Panels a-c: Discrete uniform distributions. Panels d-f: Continuous uniform distributions. Panels g-i: Gamma distributions. In all simulations, the percept e_I was sampled from a normal distribution with a mean equal to the stimulus d and a standard deviation σ_I of 1.

Fig 4 illustrates the effect of the number of possible values for |d|, assuming a finite number of possible values as well as an equal probability of each value. When there are only few possible values for |d|, confidence in choices with neutral evidence is below .75 (see Fig4 a, b). Only when the number of discrete possible values increases – and thus the distribution which |d| is sampled from becomes more like a continuous uniform distribution - confidence in choices about neutral evidence becomes close to .75 (see Fig 4c, d).

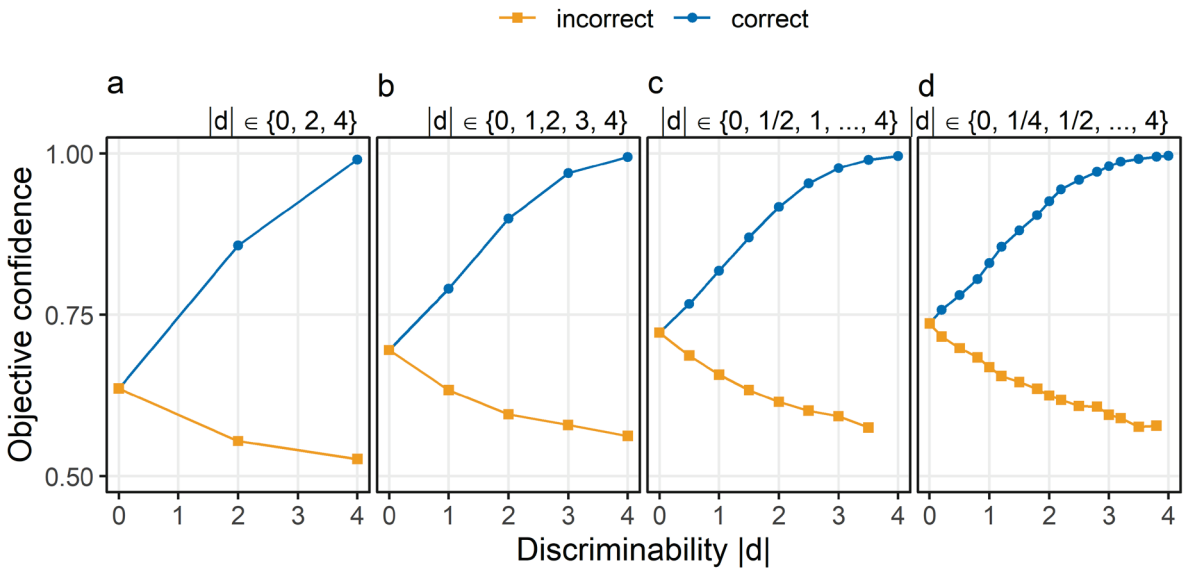

Fig 4. Objective confidence in standard model depending on the number of different levels of discriminability. Confidence (y-axis) is shown as a function of discriminability (x-axis) in correct choices (blue) and incorrect choices (orange). Different panels show different numbers of levels of discriminability |d|, sampled from discrete uniform distributions. Possible values of |d| are 0, 2, and 4 (Panel a), 0, 1, 2, 3, and 4 (Panel b), 0, ½, 1, ..., 4 (Panel c) or 0, ¼, ½, ..., 4 (Panel d). The percept e_I was sampled from a normal distribution with a mean equal to the stimulus d and a standard deviation σ_I of 1.

Fig 5 shows the statistical signatures of confidence assuming only two equally probable values of |d|. In this case, confidence in choices about neutral evidence is not .75, irrespective of whether neutral stimuli are paired with hard decisions (see Fig 5a, b), or even with more easy decisions (see Fig 5c, d).

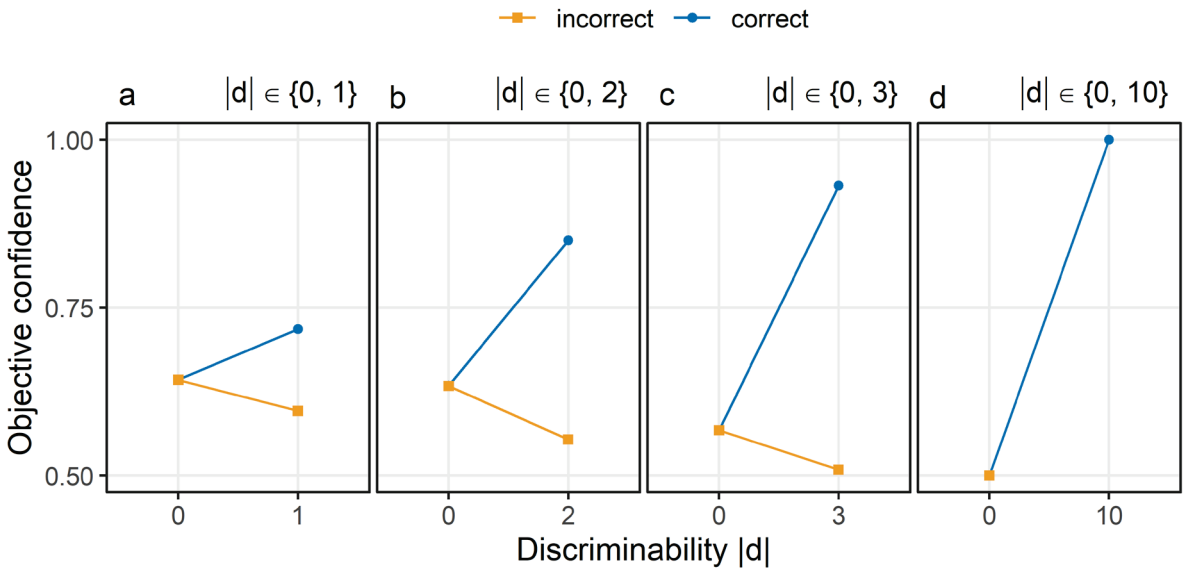

Fig 5. Objective confidence in the standard model if discriminability is either 0 or maximal. Discriminability |d| was always sampled from discrete uniform distributions with only two values. One of the two possible values was always 0, indicating neutral stimuli with respect to the choice options. The second possible value of |d| was 1 (Panel a), 2 (Panel b), 3 (Panel c), or 10 (Panel d). Confidence (y-axis) is shown as a function of discriminability (x-axis) in correct choices (blue) and incorrect choices (orange). The percept e_I was sampled from a normal distribution with a mean equal to the stimulus d and a standard deviation σ_I of 1.

General model

What is the signature of confidence expected from the general model? As can be seen from Fig 6, the general model is compatible with both the folded X-signature and the double increase signature. When σ_d is small and thus the evidence about discriminability is reliable (see Fig 6, left column), confidence approaches .5 when discriminability is 0. In addition, confidence increases with discriminability for both in correct choices well as in incorrect choices, i.e. confidence is characterised by what we refer to as the double increase signature. These patterns are the same across different distributions of discriminability (Fig 6, different rows). When σ_d is large and thus there is only corrupted evidence about discriminability (see Fig 6, right column), the pattern of confidence is the same as for the standard model (cf. Fig 3). When σ_d increases (see Fig 6, central columns), confidence in choices about stimuli with $d = 0$ increases. Additionally, when σ_d increases, the correlation between discriminability and confidence in incorrect choices becomes more negative, eventually switching sign from positive to negative.

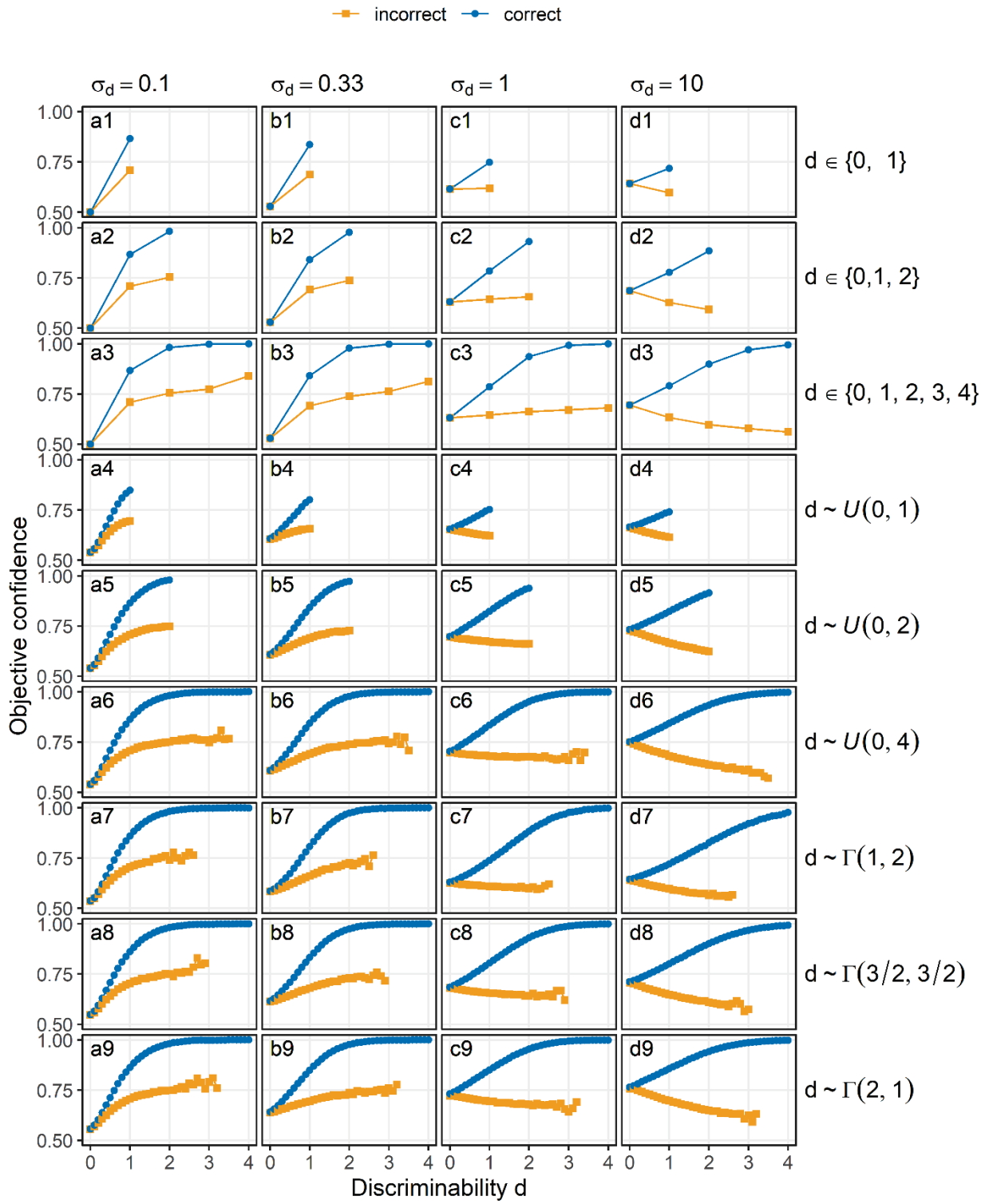

Fig 6. Objective confidence according to the general model of confidence. The sensory evidence about the identity e_I and evidence about discriminability e_d were both sampled from normal distributions, with standard deviations $\sigma_I = 1$ and σ_d varying across columns. Confidence (y-axis) is shown as a function of discriminability (x-axis) in correct trials (blue) and incorrect trials (orange). Panels a1-a9: $\sigma_d = 0.1$. Panels b1-b9: $\sigma_d = 0.33$. Panels c1-c9: $\sigma_d = 1$. Panels d1-d9: $\sigma_d = 10$. Different rows indicate different distributions of

discriminability within the simulated experiments. Rows 1-3: Discrete uniform distributions, rows 4-6: continuous uniform distributions, rows 7-9: Gamma distributions.

Accuracy of confidence

Accuracy of confidence was assessed by the information entropy of choice accuracy conditioned on confidence $H(A|c)$. The information entropy is a measure of prediction error motivated by the free energy principle [38]: $H(A|c)$ reflects the uncertainty with respect to choice accuracy given confidence; if choice accuracy is perfectly specified by confidence, $H(A|c)$ is zero. Fig 7 compares $H(A|c)$ between confidence based on evidence about the identity e_I only and confidence based on evidence about the identity e_I and evidence about discriminability e_d. The assumption that confidence is based exclusively on evidence about e_I is equivalent to the standard model. Fig 7 shows that when the standard deviation of the evidence about discriminability σ_d is low, confidence based on e_I and e_d is associated with a lower information entropy of accuracy conditioned on confidence than confidence based on e_I alone. This means that when there is an accurate estimate of discriminability, confidence that takes the evidence about discriminability into account is associated with a smaller prediction error than confidence ignoring evidence about discriminability. For larger values of σ_d, $H(A|c)$ is the same between confidence based on e_I and e_d and confidence based on e_I, meaning that there is no longer a benefit of the estimate of discriminability when the estimate was too noisy. Importantly, even when σ_d is very large, there is never a case when confidence based on e_I and e_d is worse than confidence based solely on e_I.

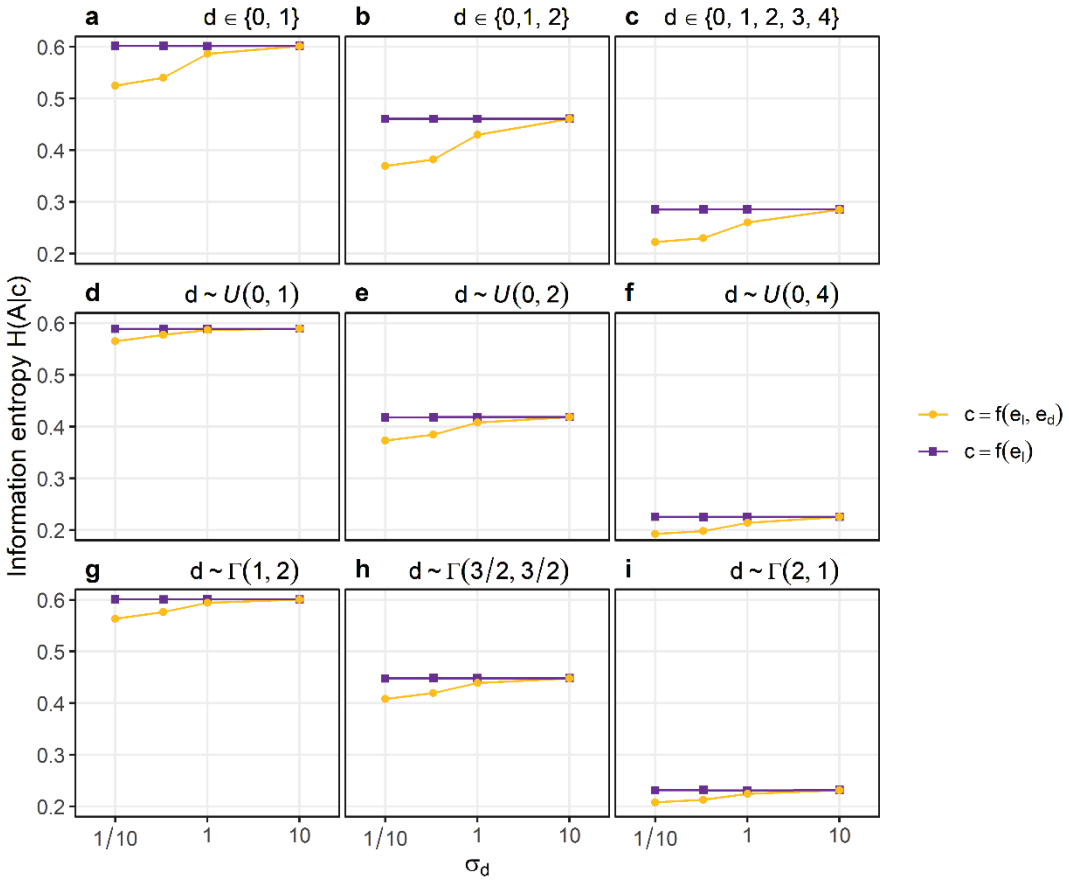

Fig 7. The information entropy of choice accuracy conditioned on confidence H(A|c). The noise parameters of the estimate of discriminability σ_d is displayed on the x-axis. Different panels indicate different distributions of discriminability within the simulated experiments. Panels a-c: Discrete uniform distributions. Panels d-f: Continuous uniform distributions. Panels g-i: Gamma distributions. Violet symbols indicate H(A|c) when confidence is calculated exclusively based on sensory evidence about the identity of the stimulus e_I. Orange symbols indicate H(A|c) when confidence is calculated based on evidence about the identity of the stimulus e_I and on evidence about discriminability e_d. The standard deviation of evidence about the identity σ_I was set to 1.

Discussion

The present study showed that the objective calculation of confidence does often not imply the folded X-signature. Even when the standard model of confidence is assumed, confidence in choices about neutral stimuli is not .75 unless discriminability is sampled from a continuous uniform distribution with high maximal discriminability. When there is sufficient evidence about discriminability as predicted by the general model, the correlation between discriminability and confidence in incorrect trials is positive, not negative. We also

showed by simulations that if observers make optimal use of the evidence, and if evidence about discriminability is available, then confidence depends on evidence about discriminability.

The observation that the Bayesian calculation of confidence does not always imply the folded X-signature corroborates the results of a previous study [26]. Adler and Ma showed that the folded X-signature depends on the distribution from which the stimulus is sampled. Specifically, confidence in incorrect choices no longer decreases with discriminability if stimuli are only probabilistically related to which choice observers ought to make. Likewise, confidence in neutral events is .75 only if the width of the stimulus distribution is quite large compared to the noise in perception. The present study shows that there are at least two more cases where confidence is not expected to follow the folded X-signature. First, when discriminability does not vary continuously but in a small number of discrete steps, optimal confidence in choices about neutral events is not .75. Notably, previous studies assuming the folded X-signature typically relied on discrete manipulations of discriminability. Second, when observers can perceive or infer discriminability on a single trial level with sufficient accuracy, objective confidence follows the double increase signature.

In summary, these observations imply that blind reliance on the folded X-signature potentially leads to false conclusions. Identifying correlates of confidence by a priori presupposing the folded X-signature is not advisable because objective confidence may not show the expected properties. Likewise, it is also not advisable to infer the computational principles underlying observed confidence judgments based on statistical signatures alone, because various different models are able to recreate the folded X-signature [26,28,29,32], just as the double increase signature [26,32,39]. Importantly, both the folded X-signature and the double increase signature are compatible with Bayesian computation of confidence,

which is why model fitting is necessary to ascertain which model is the generative model of the data [26].

Why should sensory evidence parallel to the choice improve objective confidence?

The double increase signature has been regarded as indicative of a suboptimal mental shortcut to the probability of being correct [33], i.e. a heuristic [36,37]. However, as evidence about discriminability in fact decreases the prediction error of confidence, the double increase signature may in some cases indicate optimal, not suboptimal calculation of confidence.

Too see why it is necessary to include e_d in the calculation of objective confidence, we can look at the formula of posterior probability of the identity according to the general model (see S2 Appendix for the derivation):

$$p(I = 1|(e_I, e_d)) = \frac{\sum_k p(d_k) \times p(e_I|d_k, I = 1) \times p(e_d|d_k)}{\sum_{j,k} p(d_k) \times p(e_I|d_k, I = j) \times p(e_d|d_k)} \quad (1)$$

In formula (1), I represents the identity of the stimulus, e_I the evidence about the identity, d discriminability, and e_d the is the evidence about discriminability. As can be seen from the formula, evidence about the discriminability e_d is needed to calculate the objective posterior probability given the evidence. This means if observers make optimal use of the evidence, and if evidence about the discriminability e_d is available, e_d ought to be included into the calculation of the posterior probability of the identity and hence confidence.

Now, to get some intuition why it is optimal to include e_d in the calculation of confidence, let us look at formula (1) more closely. The Bayesian computation of the posterior probability divides the likelihood of the evidence about the identity e_I given the identity 1 (the term in the numerator) by the sum of the likelihood of e_I given I = -1 and the likelihood of e_I given I = 1 (in the denominator). Calculating the likelihood of e_I requires knowledge of the distribution from which e_I is sampled. However, according to the model,

e_I is sampled from a Gaussian whose mean not only depends on I, but also on d. For this reason, the likelihood of e_I given I is calculated by multiplying the prior probability of a specific level discriminability $p(d_k)$ with the likelihood of e_I given the level discriminability and the identity $p(e_I|d_k, I)$, and summing these terms across all levels of discriminability. Conceptually, these terms imply a consideration how plausible e_I is given the identity and given the level of discriminability, weighted by the plausibility of that level of discriminability. These terms are summed over all possible values of discriminability. The product of $p(d_k)$ and $p(e_I|d_k, I)$ represents the case of the standard model: Observers know how plausible each degree of discriminability is across the experiment, and based on that prior information, they evaluate the plausibility of e_I. The novel feature of the general model is the inclusion of the probability of evidence about discriminability given discriminability $p(e_d|d_k)$. Conceptually, $p(e_d|d_k)$ implies a consideration how plausible the level of discriminability based on the evidence about the discriminability. As can be seen in the formula, $p(e_d|d_k)$ is multiplied with $p(d_k)$ and $p(e_I|d_k, I)$. Thus, in the general model, observers attach weight to $p(e_I|d_k, I)$ not only based the prior knowledge of the distribution of discriminability within the experiment, but they also evaluate the plausibility of each degree of discriminability based on sensory evidence about the discriminability. Thus, evidence about the discriminability improves the efficiency of the evaluation of e_I because evaluating the plausibility of $p(e_I|I)$ requires knowledge about d, and some additional information about the discriminability is better than the prior distribution alone. If $p(e_d|d_k)$ is the same across all levels of discriminability, the general model makes the same predictions as the standard model; conceptually, identical $p(e_d|d_k)$ across all levels of discriminability represents the case when there is no information about discriminability on a single trial basis.

Empirical support for folded X- and the double increase signature

What is the empirical evidence about these hypothesized signatures of confidence? Several previous experiments were indeed in accordance with the folded X-signature of confidence. In an auditory discrimination task [14], a general knowledge task [14], as well as a visual two-alternative forced choice tasks [40], confidence increased with discriminability in correct trials, decreased with discriminability in incorrect trials, and was medium when stimuli could not be distinguished. The folded X-signature was also consistent with rats' willingness to wait for reward in an odour discrimination task [7,24], which can be seen as a marker of confidence in non-humans.

However, six other studies based on human observers were not consistent with the folded X-signature, and three of these studies revealed the double increase pattern instead. In two random dot motion discrimination tasks, coherence of motion was positively, not negatively, associated with confidence in incorrect trials [41,42]. Likewise, in a masked orientation discrimination task, confidence in incorrect trials increased with stimulus-onset-asynchrony as well [32]. Two studies revealed a relationship between confidence in incorrect trials and discriminability that was essentially flat. In a second masked orientation discrimination task, in which observers' confidence was assessed by asking observers on which of two subsequent orientation judgments they were willing to bet, confidence in incorrect trials was approximately constant across levels of stimulus contrast [43]. Moreover, in a low-contrast orientation discrimination task, the average confidence in incorrect trials was approximately constant across task difficulty levels [44]. Finally, in a discrimination task about the average orientation of a sequence of oriented Gabor patches, one subset of observers showed the folded X-signature and another subset the double increase signature [33], although the interpretation of the inverse variability of sequence of oriented Gabor patches as discriminability is controversial [26].

Overall, these studies suggested that the folded X-signature is by no means universal. Although there is empirical support for the folded X-signature in some experiments, in other experiments the pattern is just opposite to what has been considered as the signature of confidence.

How can the differences between those studies be explained? One possibility is that some experimental tasks allow observers to estimate the discriminability on a single trial basis, as predicted by the general model: Strikingly, all studies that reported an increase of confidence and incorrect choices with discriminability were based on psychophysical tasks where the stimulus was composed out of one feature that defined the response as well as an orthogonal manipulation of discriminability: In the random dot motion discrimination tasks, participants responded to the direction of motion, and the discriminability was manipulated by the coherence of the motion signal [41,42]. Likewise, in the masked orientation task, the identity of the stimulus was defined by the orientation of the stimulus, while discriminability was manipulated by the time between stimulus onset and mask onset [32]. In contrast, those studies that observed that confidence in incorrect choices decreased with discriminability all aimed to vary the evidence more directly by using stimulus material providing different mixtures of evidence to the observer: The auditory discrimination experiment delivered click streams to both ears of the observers, and participants had to indicate which click rate was faster. Importantly, evidence was varied by the ratio between click frequencies in the two streams [14]. Likewise, the general knowledge task required observers to decide which of two countries had a greater population, with discriminability defined as the log ratio of the population size of the two countries [14]. Finally, participants in one of the two visual two-alternative forced choice tasks indicated which of two presented textured stimuli showed had un unequal amount of white and black squares. The difficulty of the task was varied by the proportion of white to black squares [40]. In all these tasks, the stimulus consisting of

mixtures of evidence about the identity might make it more difficult to estimate discriminability.

An alternative explanation for the differences between studies relying on the timing of the confidence measurement is not consistent with all the existing studies. It has been argued that asking observers to indicate their choice and their confidence at the same time interferes with the confidence report [14]. For example, asking participants to report confidence and choice at the same time might be sufficient to induce a report strategy that is no longer based on posterior probabilities, but on heuristics [45]. Additionally, measuring confidence after the choice may allow observers to collect additional evidence after the choice or even change their minds [3,40,41,46,47]. In favour of the timing-based explanation, those studies to report a decrease of confidence with discriminability assessed first the choice and confidence only after the choice [14,40]. The studies to report the opposite pattern more often recorded confidence simultaneously with the response [41,42]. Nevertheless, at least in the masked orientation discrimination task, the timing of the responses does not provide a satisfying explanation, because an increase of confidence in incorrect choices with discriminability was consistently observed irrespective of whether confidence was assessed at the same time as the choice or afterwards [32]. Future experiments appear necessary to test if the timing of the confidence measurement influences signatures of confidence in the other experimental paradigms.

Is there other empirical support for the hypothesis that confidence is not only based on sensory evidence about the identity of the stimulus, but also on evidence about discriminability? There is evidence that the brain represents estimates of discriminability: A recent neuro-imaging study showed that neural areas in posterior parietal cortex and ventral striatum track sensory reliability independently of the choice [4]. To our knowledge, only one study so far included evidence about discriminability into a formal modelling analysis. In

a masked orientation discrimination task, confidence was best explained by a combination of evidence about the identity of the stimulus as well as the general visibility of the stimulus, although the study did not test whether evidence about the identity of the stimulus and visibility were combined in a Bayesian fashion [32]. In contrast, when the double increase signature was observed in random dot kinematograms, the increase of confidence in errors with discriminability was explained by an influence of decision times of confidence [41,42]. However, at least in the masked orientation discrimination task, decision times cannot not account for the increase of confidence in errors with discriminability because decision time in incorrect trials was uncorrelated with discriminability [32].

Although more experiments are clearly necessary to investigate the relationship between confidence and decision time, the hypothesis regarding e_d gains some plausibility due to converging evidence that human confidence is informed by many cues. One mechanism may rely on the variability of e_I: In a random dot motion discrimination task, confidence depended on the consistency of the random dot motion, although discrimination performance was equated [48]. Additionally, when observers discriminated the average colour of an array of coloured shapes, confidence was not only determined by the distance of the average colour to the category boundary, but was also affected by the variability of colour across the array [49]. A second mechanism may rely on the elapsed time during decision making: In in a global motion discrimination task, the time required to make a decision was varied while the sensory evidence about the motion direction was equated, showing that decision time directly informed confidence [41]. Given that human metacognition appears to make use of such a variety of cues, it seems plausible to us that sensory evidence about discriminability may be involved as well.

Conclusion

To summarize, the present paper argues that previously postulated signatures of confidence can be misleading. On theoretical grounds, it can be expected that in many psychophysical tasks, confidence in incorrect choices increases, not decreases with discriminability. On empirical grounds, it must be acknowledged that statistical signatures of confidence can only be observed for some tasks, while it does hold true for other tasks. Overall, it is not legitimate to identify neural correlates of confidence by assuming a specific signature of confidence a priori. When statistical properties are used to track correlates of confidence, it appears essential to empirically assess the pattern of confidence in each single task using behavioural markers of confidence.

Material and Methods

All simulations were conducted using the free software R [50]. Each simulated experiment consisted of 4×10^6 trials.

Standard model

For the standard model, three sets of simulations were performed. Each simulation started with sampling the stimulus d for each single trial of the simulated experiment. We assumed that the identity of the stimulus was -1 and 1 for 2×10^6 trials each. Then, we sampled discriminability |d|. For the first set of simulations, we simulated 9 experiments, where the discriminability |d| was sampled from a different distribution for each of the nine experiments:

- discrete uniform distribution with the possible values 0, and 1
- discrete uniform distribution with the possible values 0, 1, and 2
- discrete uniform distribution with the possible values 0, 1, 2, 3, and 4
- continuous uniform distribution with min = 0 and max = 1
- continuous uniform distribution with min = 0 and max = 2
- continuous uniform distribution with min = 0 and max = 4
- gamma distribution with a shape $\alpha = 1$ and rate $\beta = 2$
- gamma distribution with a shape $\alpha = 1.5$ and rate $\beta = 1.5$
- gamma distribution with a shape $\alpha = 2$ and rate $\beta = 1$.

The parameters of the gamma distribution were chosen so that the mean and variance of the distribution matched the discrete uniform distributions.

The second set of simulations with the standard model involved four simulated experiments.|d| was always sampled from a discrete uniform distribution, but we varied the set from which |d| was sampled:

- Possible values were 0, 2, and 4
- Possible values were 0, 1, 2, 3, and 4
- Possible values were 0, ½, 1, ..., 4
- Possible values were 0, ¼, ½, ..., 4

For the third set of simulations with the standard model, $|d|$ was again always sampled from a discrete uniform distribution. In each of the 4 simulated experiments, there were only two possible values of $|d|$, one of which was always 0. The other possible value of $|d|$ were 1, 2, 3, and 10, respectively.

Then, for each single trial of the stimulated experiments, the sensory evidence e_I was sampled from Gaussian distributions with $M = d$ and $\sigma_I = 1$. The choice ϑ was -1 if $e_I < 0$ and 1 otherwise. The accuracy of the choice was defined as correct when I and ϑ were the same. For each single trial, the posterior probability of a correct choice given the percept and the choice $p(A=1|\vartheta, e_I)$ was calculated using the formulae S1 Appendix.

General model

For the simulation based on the general model, we simulated 36 experiments, one for each combination of 9 possible distributions from which the discriminability d was drawn, and 4 possible levels of noise σ_d with respect to the sensory evidence e_d about the discriminability. In each experiment, we first sampled the identity of the stimulus $I \in \{-1,1\}$ for each single trial of the experiment. It was assumed that both identities of the stimulus $I = \{-1,1\}$ occurred 2×10^6 times. Then, the discriminability d was drawn for each trial of the experiment. We used the same distributions as in the first set of simulations for the standard model. Then, for each single trial of the experiment, the evidence about the identity of the stimulus e_I was sampled from Gaussian distributions with $M = d \times I$ and $\sigma = 1$. When e_I was greater than zero, observers were assumed to make the choice $\vartheta = 1$, and $\vartheta = -1$

otherwise. When the choice matched the identity of the stimulus, the choice was considered correct. The evidence about discriminability e_d was sampled from Gaussian distributions with $M = d$ and the standard deviation of σ_d. σ_d varied across experiments with the possible values 1/10, 1/3, 1, and 10.

Finally, confidence c was calculated for each single trial as the posterior probability of a correct choice given the sensory evidence for identity, sensory evidence for discriminability, and choice $p(A=1|\vartheta, e_d, e_I)$ was calculated using the formulae S2 Appendix.

Accuracy of confidence

The information entropy of choice accuracy conditioned on confidence $H(A|c)$ can be calculated as

$$H(A|c) = -\frac{1}{n} \times \sum_{j} \left(\log(A_j \times c_j + (1 - A_j) \times (1 - c_j))\right) \quad (3)$$

where n is the number of simulated trials, A_j is the accuracy in trial j, and c_j is the confidence in trial j.

Acknowledgements

We are grateful to Christina Linner and Florian Sprang for helpful comments on a previous version of this paper.

References

1. Boldt A, Yeung N. Shared Neural Markers of Decision Confidence and Error Detection. J Neurosci. 2015;35: 3478–3484. doi:10.1523/JNEUROSCI.0797-14.2015
2. Faivre N, Filevich E, Solovey G, Kühn S, Blanke O. Behavioural, modeling, and electrophysiological evidence for supramodality in human metacognition. J Neurosci. 2018;38: 0322–17. doi:10.1523/JNEUROSCI.0322-17.2017
3. Fleming SM, van der Putten EJ, Daw ND. Neural mediators of changes of mind about perceptual decisions. Nat Neurosci. 2018;21: 617–624. doi:10.1038/s41593-018-0104-6
4. Bang D, Fleming SM. Distinct encoding of decision confidence in human medial prefrontal cortex. Proc Natl Acad Sci. 2018;115: 201800795. doi:10.1073/pnas.1800795115
5. Hebart MN, Schriever Y, Donner TH, Haynes JD. The Relationship between Perceptual Decision Variables and Confidence in the Human Brain. Cereb Cortex. 2016;26: 118–130. doi:10.1093/cercor/bhu181
6. Peters MAK, Thesen T, Ko YD, Maniscalco B, Carlson C, Davidson M, et al. Perceptual confidence neglects decision-incongruent evidence in the brain. Nat Hum Behav. 2017;1: 1–21. doi:10.1038/s41562-017-0139
7. Kepecs A, Uchida N, Zariwala H, Mainen ZF. Neural correlates, computation and behavioural impact of decision confidence. Nature. 2008;455: 227–231. doi:10.1038/nature07200
8. Komura Y, Nikkuni A, Hirashima N, Uetake T, Miyamoto A. Responses of pulvinar neurons reflect a subject ' s confidence in visual categorization. Nat Neurosci. Nature Publishing Group; 2013;16: 749–755. doi:10.1038/nn.3393
9. Fetsch CR, Kiani R, Newsome WT, Shadlen MN. Effects of Cortical Microstimulation

on Confidence in a Perceptual Decision. Neuron. Elsevier; 2014;83: 797–804. doi:10.1016/j.neuron.2014.07.011

10. Lak A, Costa GM, Romberg E, Koulakov AA, Mainen ZF, Kepecs A. Orbitofrontal Cortex Is Required for Optimal Waiting Based on Decision Confidence. Neuron. Elsevier Inc.; 2014;84: 190–201. doi:10.1016/j.neuron.2014.08.039
11. Nakamura N, Watanabe S, Betsuyaku T, Fujita K. Do birds (pigeons and bantams) know how confident they are of their perceptual decisions? Anim Cogn. 2011;14: 83–93. doi:10.1007/s10071-010-0345-6
12. Kiani R, Shadlen MN. Representation of confidence associated with a decision by neurons in the parietal cortex. Science (80-). 2009;324: 759–764. doi:10.1126/science.1169405
13. Odegaard B, Grimaldi P, Cho SH, Peters MAK, Lau H, Basso MA. Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. Proc Natl Acad Sci. 2018; 201711628. doi:10.1073/pnas.1711628115
14. Sanders JI, Hangya B, Kepecs A. Signatures of a Statistical Computation in the Human Sense of Confidence. Neuron. 2016;90: 499–506. doi:10.1016/j.neuron.2016.03.025
15. Hangya B, Sanders JI, Kepecs A. A Mathematical Framework for Statistical Decision Confidence. Neural Comput. 2016;28: 1840–1858. doi:10.1162/NECO_a_00864
16. Pouget A, Drugowitsch J, Kepecs A. Confidence and certainty: distinct probabilistic quantities for different goals. Nat Neurosci. 2016;19: 366–374. doi:10.1038/nn.4240
17. Cox RT. Probability, Frequency and Reasonable Expectation. Am J Phys. 1946;14: 1–13.
18. Jaynes ET. Probability Theory: The Logic of Science. G. Larry Bretthorst, editor. Cambridge University Press; 2003.

19. Meyniel F, Sigman M, Mainen ZF. Perspective Confidence as Bayesian Probability: From Neural Origins to Behavior. Neuron. Elsevier Inc.; 2015;88: 78–92. doi:10.1016/j.neuron.2015.09.039

20. Kepecs A, Mainen ZF. A computational framework for the study of confidence in humans and animals. Philos Trans R Soc B. 2012;367: 1322–1337. doi:10.1098/rstb.2012.0037

21. Drugowitsch J. Becoming Confident in the Statistical Nature of Human Confidence Judgments. Neuron. Elsevier Inc.; 2016;90: 425–427. doi:10.1016/j.neuron.2016.04.023

22. Urai AE, Braun A, Donner TH. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. Nat Commun. 2017;8: 14637. doi:10.1038/ncomms14637

23. Braun A, Urai AE, Donner TH. Adaptive History Biases Result from Confidence-weighted Accumulation of Past Choices. J Neurosci. 2018;38: 2189–17. doi:10.1523/JNEUROSCI.2189-17.2017

24. Lak A, Nomoto K, Keramati M, Sakagami M, Kepecs A. Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision. Curr Biol. Elsevier Ltd.; 2017;27: 821–832. doi:10.1016/j.cub.2017.02.026

25. Rolls ET, Grabenhorst F, Deco G. Decision-Making, Errors, and Confidence in the Brain. J Neurophysiol. 2010;104: 2359–2374. doi:10.1152/jn.00571.2010

26. Adler WT, Ma WJ. Limitations of proposed signatures of Bayesian confidence. Neural Comput. 2018; 218222. doi:10.1162/neco_a_01141

27. Adler WT, Ma WJ. Comparing Bayesian and non-Bayesian accounts of human confidence reports. PLOS Comput Biol. 2018;14: e1006572. doi:10.1371/journal.pcbi.1006572

28. Fleming SM, Daw ND. Self-evaluation of decision-making: A general Bayesian

framework for metacognitive computation. Psychol Rev. 2017;124: 91–114.

29. Insabato A, Pannunzi M, Deco G. Neural correlates of metacognition: A critical perspective on current tasks. Neurosci Biobehav Rev. 2016;71: 167–175. Available: http://www.mendeley.com/research/no-title-avail/

30. Xu Y. The Neural Fate of Task-Irrelevant Features in Object-Based processing. J Neurosci. 2010;30: 14020–14028. doi:10.1523/JNEUROSCI.3011-10.2010

31. Marshall L, Bays P. Obligatory encoding of task-irrelevant features depletes working memory resources. J Vis. 2013;12: 853–853. doi:10.1167/12.9.853

32. Rausch M, Hellmann S, Zehetleitner M. Confidence in masked orientation judgments is informed by both evidence and visibility. Attention, Perception, Psychophys. 2018;80: 134–154. doi:10.3758/s13414-017-1431-5

33. Navajas J, Hindocha C, Foda H, Keramati M, Latham PE, Bahrami B. The idiosyncratic nature of confidence. Nat Hum Behav. Springer US; 2017;1: 810–818. doi:10.1038/s41562-017-0215-1

34. Zylberberg A, Barttfeld P, Sigman M. The construction of confidence in a perceptual decision. Front Integr Neurosci. 2012;6: 1–10. doi:10.3389/fnint.2012.00079

35. Samaha J, Switzky M, Postle BR. Confidence boosts serial dependence in orientation estimation. J Vis. 2019;19: 25. doi:10.1167/19.4.25

36. Gigerenzer G, Gaissmaier W. Heuristic Decision Making. Annu Rev Psychol. 2011;62: 451–482. doi:10.1146/annurev-psych-120709-145346

37. Rahnev D, Denison RN. Suboptimality in Perceptual Decision Making. Behav Brain Sci. 2018;41: e223. doi:10.1017/S0140525X18000936

38. Friston K. The free-energy principle: A unified brain theory ? Nat Rev Neurosci. 2010;11: 127–138. doi:10.1038/nrn2787

39. Rausch M, Zehetleitner M. Modelling visibility judgments using models of decision

confidence. In: PsyArXiv. 2019. doi:10.31219/osf.io/7dakz

40. Moran R, Teodorescu AR, Usher M. Post choice information integration as a causal determinant of confidence: Novel data and a computational account. Cogn Psychol. Elsevier Inc.; 2015;78: 99–147. doi:10.1016/j.cogpsych.2015.01.002

41. Kiani R, Corthell L, Shadlen MN. Choice Certainty Is Informed by Both Evidence and Decision Time. Neuron. Elsevier Inc.; 2014;84: 1329–1342. doi:10.1016/j.neuron.2014.12.015

42. van den Berg R, Anandalingam K, Zylberberg A, Kiani R, Shadlen MN, Wolpert DM. A common mechanism underlies changes of mind about decisions and confidence. Elife. 2016;5: e12192. doi:10.7554/eLife.12192

43. Peters MAK, Lau H. Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. Elife. 2015;4: e09651. doi:10.7554/eLife.09651

44. Rausch M, Zehetleitner M. Visibility is not equivalent to confidence in a low contrast orientation discrimination task. Front Psychol. 2016;7: 591. doi:10.3389/fpsyg.2016.00591

45. Aitchison L, Bang D, Bahrami B, Latham PE. Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making. PLoS Comput Biol. 2015;11: 1–23. doi:10.1371/journal.pcbi.1004519

46. Resulaj A, Kiani R, Wolpert DM, Shadlen MN. Changes of mind in decision-making. Nature. 2009;461: 263–266. doi:10.1038/nature08275

47. Pleskac TJ, Busemeyer JR. Two-Stage Dynamic Signal Detection : A Theory of Choice , Decision Time, and Confidence. Psychol Rev. 2010;117: 864–901. doi:10.1037/a0019737

48. Spence ML, Dux PE, Arnold DH. Computations Underlying Confidence in Visual

Perception. J Exp Psychol Hum Percept Perform. 2016;42: 671–682. doi:10.1037/xhp0000179

49. Boldt A, Gardelle V De, Yeung N. The Impact of Evidence Reliability on Sensitivity and Bias in Decision Confidence. J Exp Psychol Hum Percept Perform. 2017;

50. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.

S1 Appendix. Derivation of the formula of objective confidence according to the standard model

According to the standard model, it is assumed that an observer selects a choice $\vartheta \in \{-1,1\}$ about the identity $I \in \{-1,1\}$ of stimulus d. The identity equals the sign of d, which is sampled in each trial either from a discrete set or from a continuous distribution. The accuracy $A \in \{0,1\}$ of the choice is defined to be 1 if $\vartheta = I$ and 0 otherwise. Observers cannot perceive d directly; instead, observers make their choices based on the sensory evidence e_I, a noisy estimate of d.

Given the model specification, the posterior probability of being correct given the sensory evidence $p(A = 1|e_I)$ can be calculated as the posterior probability that identity is the same as the selected choice option, given the sensory evidence. In the following, we consider the case that the observer decides that the identity is 1; formulae for the decision that the identity is -1 can be derived just in the same way.

According to Bayes' rule, $p(I = 1|e_I)$ can be calculated as:

$$\overbrace{p(I = 1|e_I)}^{\text{posterior}} = \frac{\overbrace{p(I = 1)}^{\text{prior}} \times \overbrace{p(e_I|\, I = 1)}^{\text{likelihood}}}{\underbrace{p(e_I)}_{\text{normalisation constant}}} \tag{2}$$

Based on the law of total probability, the normalization constant $p(e_I)$ can be expressed as:

$$p(e_I) = \sum_j p(I = j) \times p(e_I \mid I = j) \tag{3}$$

For the purpose of the present analysis, we assumed that the two choice options are equally likely, i.e. the prior probabilities $p(I = -1)$ and $p(I = 1)$ are both 0.5. Therefore, (2) and (3) can be combined and simplified to:

$$p(I = 1|e_I) = \frac{p(e_I \mid I = 1)}{\sum_j p(e_I \mid I = j)} \quad (4)$$

If d is sampled from a finite set of n elements, the denominator of the fraction in (4) can be expressed as a sum of the likelihood of the sensory evidence conditioned on d over all possible values of d weighed by the probability of the specific d. For the numerator, the sum takes into account only positive values of d:

$$p(I = 1|e_I) = \frac{\sum_{k,d_k>0} p(d_k)\, p(e_I|d_k)}{\sum_k p(d_k)\, p(e_I|d_k)} \quad (5)$$

If d is sampled from a continuous distribution instead, numerator and denominator of the fraction in (4) can be expressed as integrals over d:

$$p(I = 1|e_I) = \frac{\int_0^\infty p(d)p(e_I|d)\, dd}{\int_{-\infty}^\infty p(d)p(e_I|d)\, dd} \quad (6)$$

In (6), *d* denotes the differential, while d denotes the stimulus.

S2 Appendix. Derivation of the formula of objective confidence according to the general model.

According to the general model, the observer is presented with a series of stimuli characterised by two features, the identity $I \in \{-1,1\}$ and discriminability d. It is assumed that observers in each trial make a choice $\vartheta \in \{-1,1\}$ about the identity $I \in \{-1,1\}$ of the stimulus. The accuracy $A \in \{0,1\}$ of the choice is 1 if $\vartheta = I$ and 0 otherwise. The decision is based on sensory evidence, which involves evidence about the stimulus strength e_d, which depends only on d, and evidence about the identity e_I, which depends on d and on I.

Given the model specification, the posterior probability of being correct given the sensory evidence $p(A = 1|(e_d, e_I))$ can be calculated as the posterior probability that the identity I of the stimulus is the same as the selected choice option, given the sensory evidence. In the following, we consider the case that the observer decides that I is 1; the formula for the choice that I is -1 can be derived analogously.

According to Bayes' rule, $p(I = 1|(e_D, e_I))$ can be calculated as:

$$\overbrace{p(I = 1|(e_I, e_d))}^{\text{posterior}} = \frac{\overbrace{p(I = 1)}^{\text{prior}} \times \overbrace{p((e_I, e_d)|I = 1)}^{\text{likelihood}}}{\underbrace{p(e_I, e_d)}_{\text{normalisation constant}}} \quad (7)$$

As we assumed again that the two choice options are equally likely and thus the prior probabilities of both identities are the same, formulae (7) can be simplified analogously to formula (6):

$$p(I = 1|(e_I, e_d)) = \frac{p((e_I, e_d)|I = 1)}{\sum_j p((e_I, e_d)|I = j)} \quad (8)$$

If d is sampled from a discrete set of elements, numerator and denominator in (8) can be expressed as a sum of likelihoods of the sensory evidence conditioned on d over the different values of d. The sum is weighed by the probability of each specific d:

$$p(I = 1|(e_I, e_d)) = \frac{\sum_k p(d_k) \times p((e_I, e_d)|(I = 1, d_k))}{\sum_{j,k} p(d_k) \times p((e_I, e_d)|(I = j, d_k))} \quad (9)$$

Given the case that d is sampled from a continuous distribution, $p(I = 1|(e_I, e_d))$ can be obtained by integration:

$$p(I = 1|e_I, e_d) = \frac{\int_0^\infty p(d) \times p((e_I, e_d)|(I = 1, d))\, dd}{\int_0^\infty \sum_j p(d) \times p((e_I, e_d)|(I = j, d))\, dd} \quad (10)$$

Again, *d* denotes the differential, while d denotes the discriminability of the stimulus. As we assume that e_I and e_d are stochastically independent when the stimulus strength d is controlled, the likelihood $p((e_I, e_d)|(I, d))$ can be calculated as:

$$p((e_I, e_d)|(I, d)) = p(e_d|d) \times p(e_I|(d, I)) \quad (11)$$

We insert formula (11) into (9) to calculate $p(I = 1|e_i, e_D)$ in the discrete case.

$$p(I = 1|(e_I, e_d)) = \frac{\sum_k p(d_k) \times p(e_d|d_k) \times p(e_I|(d_k, I = 1))}{\sum_{j,k} p(d_k) \times p(e_d|d_k) \times p(e_I|(d_k, I = j))} \quad (12)$$

Finally, we insert formula (11) into (10) to calculate $p(I = 1|(e_I, e_d))$ in the continuous case.

$$p(I = 1|e_I, e_d) = \frac{\int_0^\infty p(d) \times p(e_d|d) \times p(e_I|(d, I = 1))\, dd}{\int_0^\infty \sum_j p(d) \times p(e_d|d) \times p(e_I|(d, I = j))\, dd} \quad (13)$$