Multiple statistical signatures of human confidence

Manuel Rausch* and Michael Zehetleitner
Katholische Universität Eichstätt-Ingolstadt, Philosophisch-Pädagogische Fakultät, Allgemeine Psychologie II, Eichstätt, Germany.
* Correspondence should be addressed to Manuel Rausch (manuel.rausch@ku.de)

Abstract

The so called standard model of confidence predicts a specific statistical signature of confidence that has been widely used to objectively identify neural, behavioural, and physiological correlates of confidence. That signature states that confidence is medium when the stimulation is absent, increases with stimulus strength in correct trials, and decreases with stimulus strength in incorrect trials. Here, we show that the standard model is a special case in which evidence about stimulus strength is omitted in the objective calculation of confidence. In the general case, objective confidence follows a different statistical signature: for both correct and incorrect choices, confidence increases with stimulus strength. Recent experiments provided evidence for both types of patterns, implying that both the special and the general case occur in experiments. Thus, relying exclusively on one type of signature of confidence is problematic.

Keywords: confidence, metacognition, decision making, perception, certainty, Bayesian brain theory

Multiple signatures of confidence

Confidence is a metacognitive evaluation of decision making: Each choice can be accompanied by some degree of confidence that the choice is correct. In the cognitive sciences as well as the neurosciences, confidence has become a flourishing research topic, uncovering the underlying neural mechanisms in humans (Boldt & Yeung, 2015; Faivre, Filevich, Solovey, Kühn, & Blanke, 2018; Fleming & Dolan, 2012; Fleming, Weil, Nagy, Dolan, & Rees, 2010) as well as animals (Churchland, Kiani, & Shadlen, 2008; Fetsch, Kiani, Newsome, & Shadlen, 2014; Kepecs, Uchida, Zariwala, & Mainen, 2008; Komura, Nikkuni, Hirashima, Uetake, & Miyamoto, 2013; Lak et al., 2014; Nakamura, Watanabe, Betsuyaku, & Fujita, 2011). 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 (Hangya, Sanders, & Kepecs, 2016; Sanders, Hangya, & Kepecs, 2016). This formalization defines confidence as the belief that a decision is correct (Pouget, Drugowitsch, & Kepecs, 2016). From a Bayesian perspective, beliefs are best formalised as probabilities (Cox, 1946; Jaynes, 2003). Decision confidence in this formalization is the posterior probability of being correct given the evidence (Meyniel, Sigman, & Mainen, 2015; Pouget et al., 2016). Thus, several predictions about objective confidence have been formally derived from the model to which we subsequently refer to as the standard model of confidence (see Fig. 1a, Kepecs et al., 2008; Sanders et al., 2016): First, the average objective confidence in correct decisions increases as a function of the stimulus strength. Second, the average confidence in incorrect decisions decreases with stimulus strength. Finally, when the stimulus is neutral about the decision options, confidence is exactly 0.75. That overall pattern, which we refer to here as *left arrowhead signature*, has been dubbed the "statistical signature of confidence" (Drugowitsch, 2016; Sanders et al., 2016). As the left arrowhead signature follows objectively from the posterior probability of being correct, it has been argued that when the left arrowhead pattern is detected in another behavioural, neural, or physiological variable, that variable should be considered a correlate of confidence (Hangya et al., 2016; Kepecs et al., 2008; Sanders et al., 2016; Urai, Braun, & Donner, 2017). Thus, it has been repeatedly used to empirically identify correlates of decision confidence (Braun, Urai, & Donner, 2018; Kepecs et al., 2008; Komura et al., 2013; Lak, Nomoto, Keramati, Sakagami, & Kepecs, 2017; Peters et al., 2017; Urai et al., 2017).

Here, we argue that the left arrowhead pattern of confidence only occurs in a special case when there is no representation of stimulus strength available for the decision process. The omission of stimulus strength in the decision process is due to an overly simplified set up of stimulus and decision variables in the formal specification of the standard model of confidence. When objective confidence is calculated from a general model of confidence, which includes representations of stimulus strength, the left arrowhead signature occurs as a special case when stimulus strength is not available for the decision process. However, when there is sufficient information about stimulus strength, a different pattern of confidence emerges, which we refer to here as the *double increase signature* (see Fig. 1b). Finally, we review recent empirical evidence showing that both the left arrowhead and the double increase signature have been observed in previous empirical studies.

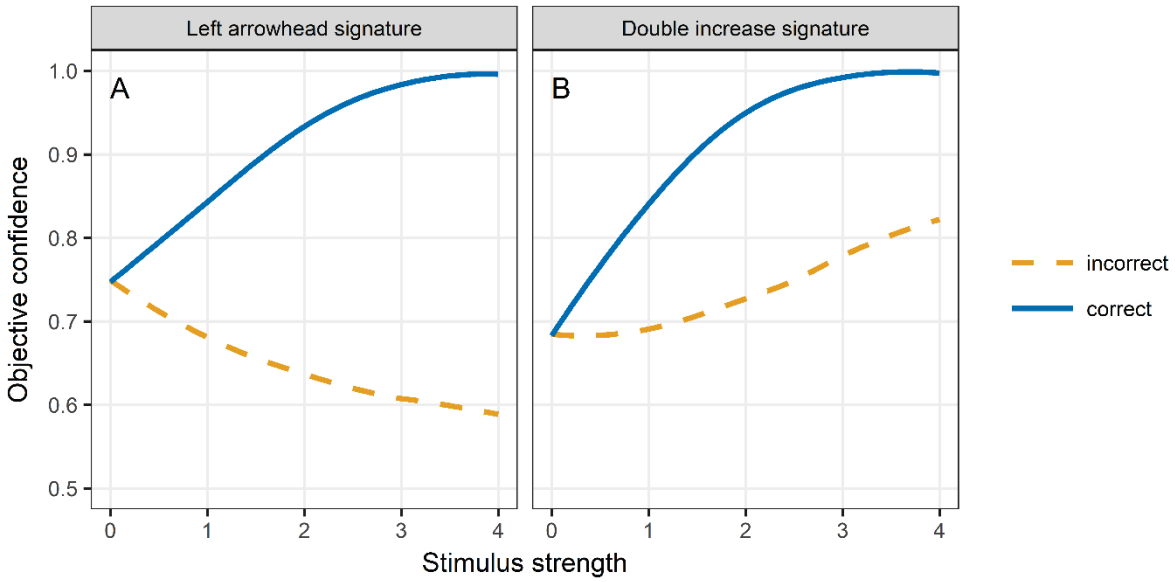

Figure 1. Objective confidence based on the posterior probability of being correct given the evidence as a function of stimulus strength (x-axis) and correctness of the decision (line types). Panel A: The left arrowhead signature, derived from the standard model of confidence (Hangya et al., 2016; Sanders et al., 2016). Panel B: The double increase signature of objective confidence derived from the general model of confidence (Rausch, Hellmann, & Zehetleitner, 2018).

The standard model of confidence

The standard model of confidence is depicted in Fig. 2. According to Sanders et al. (2016), when an observer is presented with a stimulus and asked to make a choice $\vartheta \in \{-1,1\}$ about a stimulus, the stimulation provides the observer with a certain amount of objective evidence d about which option to choose. d is a continuous variable that

differentiates between the two options of ϑ: The greater is the distance between d and 0, the easier is the choice. 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 $\vartheta = 1$. However, observers cannot perceive d directly, instead, the choice is based on noisy sensory evidence $\hat{d}$ (referred to as percept by Sanders et al.), which can be considered an estimate of d. The choice is a function of the percept $\hat{d}$. Finally, confidence c is a function of the sensory evidence $\hat{d}$ and the choice ϑ. The standard model makes no specific assumptions about the functions transferring objective stimulus evidence d to sensory evidence $\hat{d}$ or sensory evidence $\hat{d}$ to the choice ϑ (Hangya et al., 2016; Sanders et al., 2016). The most frequent approach is to model $\hat{d}$ as a random sample from a Gaussian with a mean of d, while ϑ is modelled as a deterministic function of d. In this case, the standard model is equivalent to signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002). It was argued that the left arrowhead signature holds independently of the specific functions underlying choice and confidence, which is why the left arrowhead signature has been considered to be a universal signature of confidence (Hangya et al., 2016; Kellen & Klauer, 2015; Sanders et al., 2016; Urai et al., 2017). Despite these claims, the left arrowhead signature may not be robust across different assumptions about the shape of the distribution of the percept, thus, limiting the applicability of the left arrowhead pattern as signature of confidence (Adler & Ma, 2017b; Rausch et al., 2018).

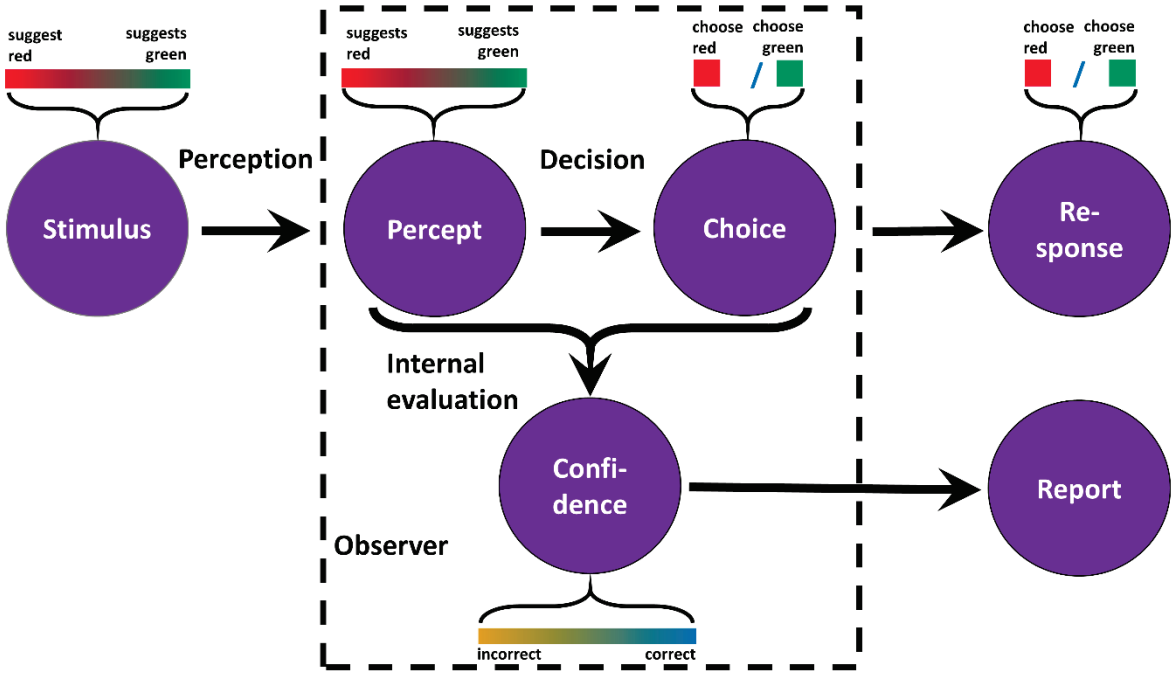

Figure 2. The standard model of confidence. The stimulus objectively supports choice options ”red” and “green” to varying degrees. As perception is noisy, the percept is a corrupted representation of the degree to which the stimulus favours a specific choice option. Confidence is the probability of making the correct choice depending on percept and choice.

The general model of confidence

The main thesis of this paper is that the objective confidence implies the left arrowhead signature only if confidence is not informed about the stimulus strength. The general model of confidence (see Fig. 3) differs from the standard model in that it differentiates between two central aspects of the stimulus: the stimulus strength d, and stimulus identity I. Stimulus strength is the variable that determines how easy/difficult the choice is. Stimulus identity is the variable that determines which choice is the correct choice. Both aspects of the stimulus are represented in the general model as sensory evidence for stimulus identity e_I, and evidence for stimulus strength e_d.

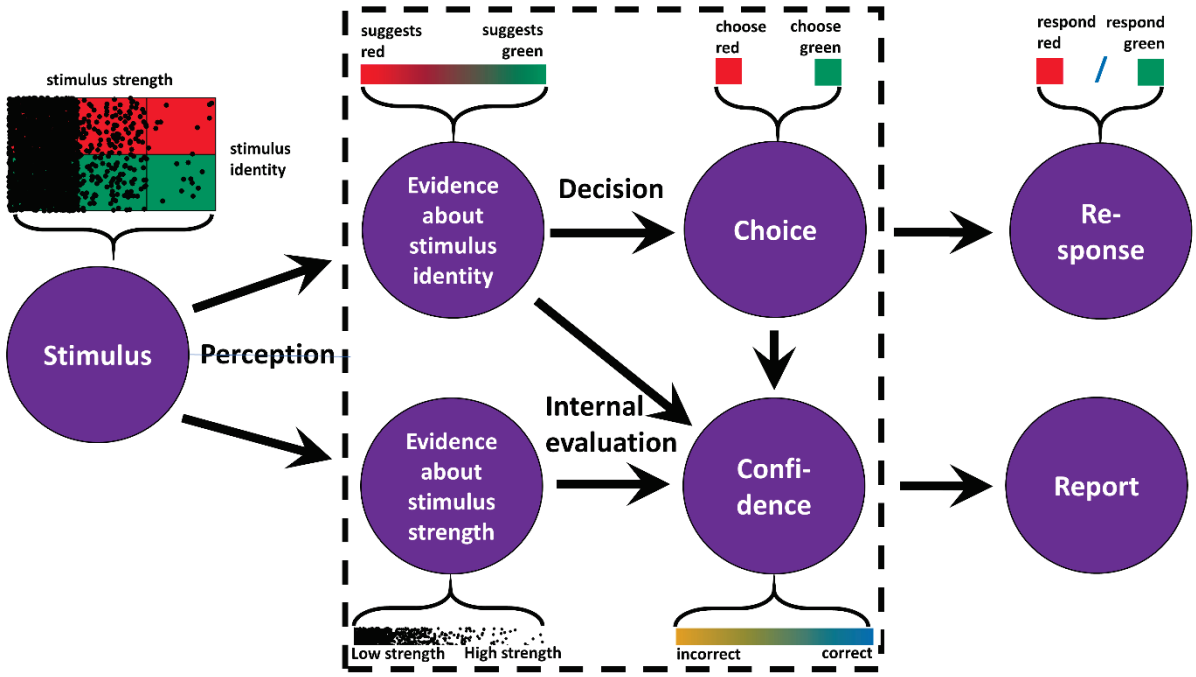

Figure 3. The general model of confidence (Rausch et al., 2018), a generalization of the standard model (Hangya et al., 2016; Sanders et al., 2016), see Peters and Lau (2015) and Navajas et al. (2017) for similar models. In many psychophysical experiments, the stimulus varies in two aspects: stimulus identity (symbolized here as red and green) and stimulus strength (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 stimulus strength, a continuous variable that differentiates between no visibility of the stimulus and full visibility. Objective confidence about the correctness of the decision is based on evidence about the identity as well as evidence about stimulus strength.

The key fact is that confidence as the posterior probability that the choice is correct given the evidence and choice $P(correct|e,\vartheta)$ is only objective if it includes all sensory evidence that is dependent on the stimulus, not only that evidence dependent on the choice-relevant features of the stimulus. That is, when decision confidence is calculated objectively, it is not exclusively based on sensory evidence for the choice options, but also on sensory evidence for the quality and reliability of perception itself in the present situation (Rausch et al., 2018). 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 their view is hindered by rain.

Analogous to traffic lights and rain, many psychophysical experiments do not manipulate the objective evidence directly; instead, two features of the stimulus are varied across the experiment. In binary choice experiments, one of the two features determines what response observers ought to give. In the general model, the response-relevant feature of the stimulus is referred to as identity $I = \{-1,1\}$. The model generates a choice ϑ about the identity I of the stimulus. For example, the stimulus could be blue or yellow, and participants need to make a choice accordingly. Choices are correct when I and ϑ are both either -1 or 1. The second feature of the stimulus is stimulus strength $d = \{d_1, d_2, \dots, d_n\}$. For instance, many experiments manipulate contrast, presentation time, or dot coherence orthogonally to stimulus identity I. It is often assumed that sensory evidence is drawn from a Gaussian distribution with mean $d \times I$. In that case, the distribution of sensory evidence e_I is centred on zero when there is no evidence in the stimulus about I. For stimuli of the identity -1, the distribution is shifted to the left to the extent determined by stimulus strength, d. For the identity 1, the distribution is shifted d signal-to-noise units to the right. However, there is not only sensory evidence about the choice-relevant stimulus feature I in the visual system, but also sensory evidence about other features of the stimulus, irrelevant to the current choice (Marshall & Bays, 2013; Xu, 2010). If sensory evidence irrelevant to the current choice is correlated with stimulus strength, it can be used as evidence about the stimulus strength e_d. Thus, objective confidence is also based on e_d. As can be seen from Fig. 4, according to the general model, confidence is more similar to the objective probability of being correct than confidence in the standard model.

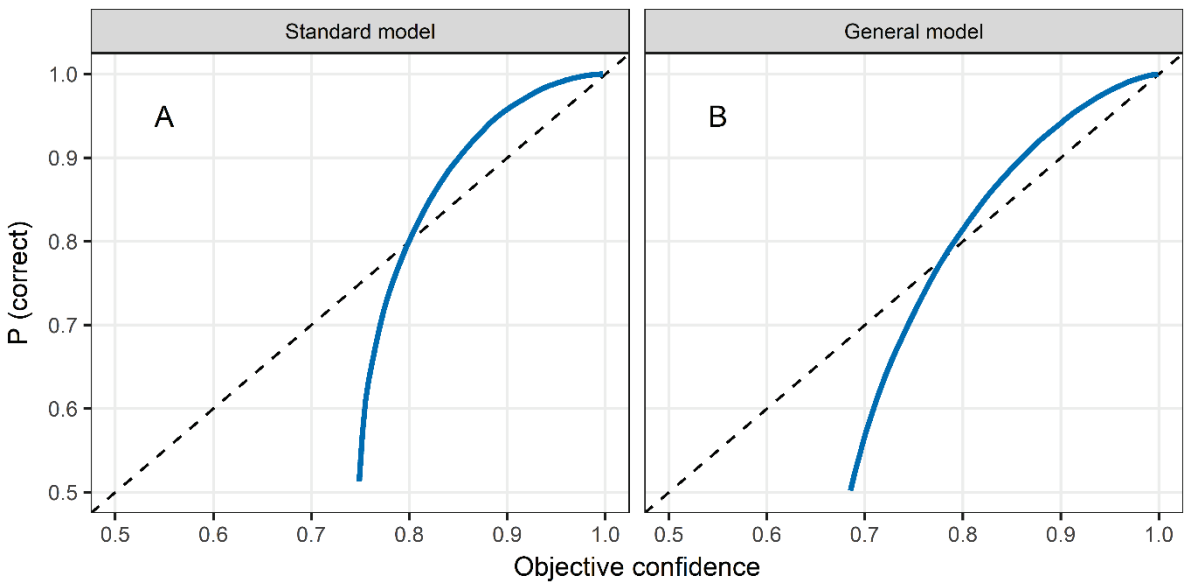

Figure 4. Relationship between the objective probability of being correct (y-Axis) and objective confidence according to the standard model (panel A) and the general model (panel B). Each location on the line represents a specific level of stimulus strength.

Too see why it is necessary to include e_d in the calculation of objective confidence, let us have a look at the formula of the posterior probability of stimulus identity I. According to Bayes' rule, the posterior probability of the identity I given the evidence is the prior probability of the identity multiplied with the probability of the evidence given the stimulus divided by the probability for the evidence. When there is both e_I and e_d, $p(I|e_I, e_d)$ is given by

$$\overbrace{p(I|e_I, e_d)}^{posterior\ I} = \frac{\overbrace{p(I)}^{prior\ I} \times \sum_d \overbrace{p(d_i)}^{prior\ d}\ \overbrace{p(e_d|d_i)}^{Likelihood\ e_d}\ \overbrace{p(e_I|I, d_i)}^{Likelihood\ e_I}}{\underbrace{p(e_I, e_d)}_{normalisation\ constant}}$$

As can be seen from the formula, the sensory evidence about the stimulus strength e_d is needed to calculate the objective posterior probability given the evidence. This means if observers make optimal use of the sensory evidence, and if sensory evidence about the stimulus strength e_d is available, e_d ought to be included into the calculation of the posterior probability of the identity and hence confidence. The posterior probability used to propose the left arrowhead pattern as signature of confidence only includes evidence about the identity e_I (Hangya et al., 2016; Sanders et al., 2016) and neglects evidence for stimulus strength e_d. However, when a representation of e_d exists, it must be included in the objective posterior.

Now, to get some intuition why it is optimal include e_d in the calculation of confidence, let us consider the posterior odds of being correct given the evidence. According to Bayes' rule, the posterior odds of being correct equals the prior odds of being correct multiplied with a likelihood ratio of the evidence given a correct and an incorrect decision:

$$\frac{P(correct|e)}{P(incorrect|e)} = \frac{P(correct)}{P(incorrect)} \times \frac{P(e|correct)}{P(e|incorrect)}$$

The prior odds reflect the observer's expectation of being correct independently of the current identity and strength of the stimulus. The optimal prior of being correct depends on stimulus strength in the experiment. If there were only strong stimuli, the prior expectation of being correct should be high and if there were only weak stimuli, the prior expectation should be low. In a binary psychophysical task with a mixture of weak and strong stimuli, we can assume that a reasonable prior probability is about 75%. The likelihood ratio reflects how strongly the present evidence indicates a correct response. When the stimulus is too difficult to discriminate, the distributions of $P(e_I|I = -1)$ and $P(e_I|I = 1)$ are the same. Hence, the average ratio of $P(e_I|correct)$ and $P(e_I|incorrect)$ will be 1. Thus, when the stimulation is absent, on average the posterior odds of being correct will be just equal to the prior odds. Now that is clearly not optimal for the observer: If a binary task involves a mixture of weak and strong stimuli, and the current stimulus is so weak that it is impossible to discern its identity, the observer still believes to be correct to 75% on average, although the true chance of being correct is only 50%. It would be more beneficial if the observers were able to adapt the prior belief of being correct from stimulus-to-stimulus based on an estimate of the stimulus strength. If observers were able to take into account that the stimulus strength had been rather weak, the prior probability would no longer be the average performance in the task, but closer to chance. Likewise, if observers were able to see that the stimulus strength had been high, they could be quite confident that the choice was correct, and thus adopt a high prior probability of being correct (Rausch et al., 2018).

What is the signature of confidence expected if observers do indeed adapt the prior probability of being correct based on sensory evidence about the stimulus strength for each single choice? As can be seen from Fig. 1b, the interaction pattern between stimulus strength and decision correctness is different from the pattern expected from the standard model: When the stimulus strength is minimal, confidence is low, not medium. In addition, confidence is expected to increase with stimulus strength both in correct decisions well as in incorrect decisions, although the increase expected in correct decision is steeper than in incorrect decisions. An increase of confidence in incorrect decisions with stimulus strength

has sometimes been regarded as indicative of a suboptimal mental shortcut to the probability of being correct (Adler & Ma, 2017a; Navajas et al., 2017), i.e. a heuristic (Gigerenzer & Gaissmaier, 2011). However, as adapting one's prior to the strength of stimulation may in fact increase the accuracy of confidence judgments, it may be misleading to characterise these effects as heuristic. Overall, there are sufficient theoretical reasons to expect that confidence does not necessarily adhere to the left arrowhead signature.

Empirical support

What is the empirical evidence about these hypothesized signatures of confidence? Several previous experiments were indeed in accordance with the left arrowhead signature of confidence. In an auditory discrimination task (Sanders et al., 2016), a general knowledge task (Sanders et al., 2016), two visual two-alternative forced choice tasks (Faivre et al., 2018; Moran, Teodorescu, & Usher, 2015), as well as an auditory two-alternative forced-choice task (Faivre et al., 2018), confidence increased with stimulus strength in correct trials, decreased with stimulus strength in incorrect trials, and was medium when stimuli could not be distinguished. The left arrowhead signature was also consistent with rats' willingness to wait for reward in an odour discrimination task (Kepecs et al., 2008; Lak et al., 2017), which can be seen as a marker of confidence in non-humans.

However, five other studies based on human observers were not consistent with the left arrowhead signature: In two random dot motion discrimination tasks, coherence of motion was positively, not negatively, associated with confidence in incorrect trials (Kiani, Corthell, & Shadlen, 2014; van den Berg et al., 2016). Moreover, in a masked orientation discrimination task, confidence in incorrect trials increased with stimulus-onset-asynchrony as well (Rausch et al., 2018). Finally, in a low-contrast orientation discrimination task (Rausch & Zehetleitner, 2016), and a discrimination task about the average orientation of a sequence of oriented Gabor patches (Navajas et al., 2017), the average confidence in incorrect trials was approximately constant across task difficulty levels. Overall, these studies suggested that the left arrowhead signature is by no means universal. Although there is empirical support for the left arrowhead 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? Strikingly, all studies that reported an increase of confidence and incorrect decisions with stimulus strength were based on psychophysical tasks where the stimulation was composed out of one feature that

defined the response as well as an orthogonal manipulation of stimulus strength: In the random dot motion discrimination tasks, participants responded to the direction of motion, and the stimulus strength was manipulated by the coherence of the motion signal (Kiani et al., 2014; van den Berg et al., 2016). Likewise, in the masked orientation task, the identity of the stimulus was defined by the orientation of the stimulus, while stimulus strength was manipulated by the time between stimulus onset and mask onset (Rausch et al., 2018). In contrast, those studies that observed that confidence in incorrect decisions decreased with stimulus strength all aimed to vary the evidence more directly by using stimulus material providing different mixtures of evidence to the observer. For example, 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 (Sanders et al., 2016). Likewise, the general knowledge task required observers to decide which of two countries had a greater population, with the objective evidence defined as the log ratio of the population size of the two countries (Sanders et al., 2016). Finally, participants in one of the two visual two-alternative forced choice task 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 (Moran et al., 2015). In all these tasks, the stimulation consisting of mixtures of evidence about the identity might make it impossible to estimate the quality of stimulation independently.

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 decision and their confidence at the same time interferes with the confidence report (Sanders et al., 2016). For example, asking participants to report confidence and the decision at the same time might be sufficient to induce a report strategy that is no longer based on posterior probabilities, but on heuristics (Aitchison, Bang, Bahrami, & Latham, 2015). Additionally, measuring confidence after the decision may allow observers to collect additional evidence after the decision or even change their minds (Kiani et al., 2014; Pleskac & Busemeyer, 2010; Resulaj, Kiani, Wolpert, & Shadlen, 2009). In favour of the timing-based explanation, those studies to report a decrease of confidence with stimulus strength recorded confidence after the decision (Moran et al., 2015; Sanders et al., 2016). The studies to report the opposite pattern more often recorded confidence simultaneously with the response (Kiani et al., 2014; van den Berg et al., 2016).

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 decisions with stimulus strength was consistently observed irrespective of whether confidence was assessed at the same time as the choice or afterwards (Rausch et al., 2018). Future experiments appear necessary to test if the timing of the confidence measurement influences signatures of confidence in the other experimental paradigms.

Is it empirically plausible that confidence is not only based on sensory evidence about the identity of the stimulus, but also on evidence about the strength of stimulation? Only one study so far addressed this possibility: In the 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, while all theories that only involved evidence about the identity fell short of accounting for the data (Rausch et al., 2018). In contrast, previous studies to observe a positive correlation between stimulus strength and confidence in incorrect decisions did not consider the possibility of evidence about stimulus strength. In those studies, the increase of confidence in errors with stimulus strength was explained by an influence of decision times of confidence (Kiani et al., 2014; van den Berg et al., 2016). However, at least in the masked orientation discrimination task, decision times cannot not account for the increase of confidence in errors with stimulus strength because decision times in incorrect trials were not affected by stimulus strength (Rausch et al., 2018).

Although more experiments are clearly necessary to investigate this issue, 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 (Spence, Dux, & Arnold, 2016). 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 (Boldt, Gardelle, & Yeung, 2017). A second mechanism may rely on the elapsed time during decision making: In in a global motion discrimination task, the time required to reach the decision was varied while the sensory evidence about the motion direction was equated, showing that decision time directly informed confidence (Kiani et al., 2014). Given that human metacognition appears to make use of such a variety of cues, it seems plausible to us that sensory evidence about the strength of stimulation may be involved as well.

Concluding remarks

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 decisions increases, not decreases with stimulus strength. 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, 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.

Acknowledgements

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

The research presented here was in parts funded by the Profor+ internal research funding of the Catholic University of Eichstätt-Ingolstadt.

References

Adler, W. T., & Ma, W. J. (2017a). Human confidence reports account for sensory uncertainty but in a non-Bayesian way. *BioRxiv*, 093203. http://doi.org/10.1101/093203

Adler, W. T., & Ma, W. J. (2017b). Limitations of proposed signatures of Bayesian confidence. *BioRxiv*, 218222. http://doi.org/10.1101/218222

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making. *PLoS Computational Biology*, *11*(10), 1–23. http://doi.org/10.1371/journal.pcbi.1004519

Boldt, A., Gardelle, V. De, & Yeung, N. (2017). The Impact of Evidence Reliability on Sensitivity and Bias in Decision Confidence. *Journal of Experimental Psychology: Human Perception and Performance*.

Boldt, A., & Yeung, N. (2015). Shared Neural Markers of Decision Confidence and Error Detection. *The Journal of Neuroscience*, *35*(8), 3478–3484. http://doi.org/10.1523/JNEUROSCI.0797-14.2015

Braun, A., Urai, A. E., & Donner, T. H. (2018). Adaptive History Biases Result from Confidence-weighted Accumulation of Past Choices. *The Journal of Neuroscience*, *38*(10), 2189–17. http://doi.org/10.1523/JNEUROSCI.2189-17.2017

Churchland, A. K., Kiani, R., & Shadlen, M. N. (2008). Decision-making with multiple alternatives, *11*(6). http://doi.org/10.1038/nn.2123

Cox, R. T. (1946). Probability, Frequency and Reasonable Expectation. *American Journal of Physics*, *14*(1), 1–13.

Drugowitsch, J. (2016). Becoming Confident in the Statistical Nature of Human Confidence Judgments. *Neuron*, *90*(3), 425–427. http://doi.org/10.1016/j.neuron.2016.04.023

Faivre, N., Filevich, E., Solovey, G., Kühn, S., & Blanke, O. (2018). Behavioural, modeling, and electrophysiological evidence for supramodality in human metacognition. *The Journal of Neuroscience*, *38*(2), 0322-17. http://doi.org/10.1523/JNEUROSCI.0322-17.2017

Fetsch, C. R., Kiani, R., Newsome, W. T., & Shadlen, M. N. (2014). Effects of Cortical Microstimulation on Confidence in a Perceptual Decision. *Neuron*, *83*(4), 797–804. http://doi.org/10.1016/j.neuron.2014.07.011

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*, 1338–1349. http://doi.org/10.1098/rstb.2011.0417

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541–1543. http://doi.org/10.1126/science.1191883.Relating

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*, 451–482. http://doi.org/10.1146/annurev-psych-120709-145346

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hangya, B., Sanders, J. I., & Kepecs, A. (2016). A Mathematical Framework for Statistical Decision Confidence. *Neural Computation*, *28*(9), 1840–1858. http://doi.org/10.1162/NECO_a_00864

Ince, D. C., Hatton, L., & Graham-Cumming, J. (2012). The case for open computer programs. *Nature*, *482*(7386), 485–488. http://doi.org/10.1038/nature10836

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. (G. Larry Bretthorst, Ed.). Cambridge University Press.

Kellen, D., & Klauer, K. C. (2015). Signal Detection and Threshold Modeling of Confidence-Rating ROCs: A Critical Test With Minimal Assumptions. *Psychological*

Review, *122*(3), 542–557.
Kepecs, A., Uchida, N., Zariwala, H., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231. http://doi.org/10.1038/nature07200
Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron*, *84*(6), 1329–1342. http://doi.org/10.1016/j.neuron.2014.12.015
Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T., & Miyamoto, A. (2013). Responses of pulvinar neurons reflect a subject ' s confidence in visual categorization. *Nature Neuroscience*, *16*(6), 749–755. http://doi.org/10.1038/nn.3393
Lak, A., Costa, G. M., Romberg, E., Koulakov, A. A., Mainen, Z. F., & Kepecs, A. (2014). Orbitofrontal Cortex Is Required for Optimal Waiting Based on Decision Confidence. *Neuron*, *84*, 190–201. http://doi.org/10.1016/j.neuron.2014.08.039
Lak, A., Nomoto, K., Keramati, M., Sakagami, M., & Kepecs, A. (2017). Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision. *Current Biology*, *27*(6), 821–832. http://doi.org/10.1016/j.cub.2017.02.026
Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory. A user's guide*. Mahwah, NY: Lawrence Erlbaum Associates.
Marshall, L., & Bays, P. (2013). Obligatory encoding of task-irrelevant features depletes working memory resources. *Journal of Vision*, *12*(9), 853–853. http://doi.org/10.1167/12.9.853
Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Perspective Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron*, *88*(1), 78–92. http://doi.org/10.1016/j.neuron.2015.09.039
Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147. http://doi.org/10.1016/j.cogpsych.2015.01.002
Morin, A., Urban, J., Adams, P. D., Foster, I., Sali5, A., Baker, D., & Sliz, P. (2012). Shining Light into Black Boxes. *Science*, *336*(6078), 159–160. http://doi.org/10.1126/science.1218263.Shining
Nakamura, N., Watanabe, S., Betsuyaku, T., & Fujita, K. (2011). Do birds (pigeons and bantams) know how confident they are of their perceptual decisions? *Animal Cognition*, *14*, 83–93. http://doi.org/10.1007/s10071-010-0345-6
Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P. E., & Bahrami, B. (2017). The idiosyncratic nature of confidence. *Nature Human Behaviour*, *1*(11), 810–818. http://doi.org/10.1038/s41562-017-0215-1
Peters, M. A. K., & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *ELife*, *4*, e09651. http://doi.org/10.7554/eLife.09651
Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., … Lau, H. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*, *1*(7), 1–21. http://doi.org/10.1038/s41562-017-0139
Pleskac, T. J., & Busemeyer, J. R. (2010). Two-Stage Dynamic Signal Detection : A Theory of Choice , Decision Time, and Confidence. *Psychological Review*, *117*(3), 864–901. http://doi.org/10.1037/a0019737
Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374. http://doi.org/10.1038/nn.4240
Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception, &*

Psychophysics, *80*(1), 134–154. http://doi.org/10.3758/s13414-017-1431-5

Rausch, M., & Zehetleitner, M. (2016). Visibility is not equivalent to confidence in a low contrast orientation discrimination task. *Frontiers in Psychology*, *7*, 591. http://doi.org/10.3389/fpsyg.2016.00591

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(September), 263–266. http://doi.org/10.1038/nature08275

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

Spence, M. L., Dux, P. E., & Arnold, D. H. (2016). Computations Underlying Confidence in Visual Perception. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(5), 671–682. http://doi.org/10.1037/xhp0000179

Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications*, *8*, 14637. http://doi.org/10.1038/ncomms14637

van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *ELife*, *5*, e12192. http://doi.org/10.7554/eLife.12192

Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.

Xu, Y. (2010). The Neural Fate of Task-Irrelevant Features in Object-Based processing. *The Journal of Neuroscience*, *30*(42), 14020–14028. http://doi.org/10.1523/JNEUROSCI.3011-10.2010