Title: Long-term perceptual priors drive confidence bias which favors prior-congruent evidence

Authors: Marika Constant[a,b], Elisa Filevich[a,b,c], Pascal Mamassian[d]

Affiliations:

[a] Humboldt-Universität zu Berlin, Faculty of Life Sciences, Department of Psychology, Unter den Linden 6, 10099 Berlin, Germany

[b] Bernstein Center for Computational Neuroscience Berlin, Philippstraße 13 Haus 6, 10115 Berlin, Germany

[c] Hector Institute for Education Sciences & Psychology, University of Tübingen, Europastraße 6, 72072, Tübingen, Germany

[d] Laboratoire des Systèmes Perceptifs, Département d'Études Cognitives, École Normale Supérieure, Paris Sciences et Lettres University, CNRS, Paris, France

Corresponding Author: Marika Constant, marika.constant@gmail.com

Abstract

Within a Bayesian framework, both our perceptual decisions and confidence about those decisions are based on the precision-weighted integration of prior expectations and incoming sensory information. This assumes priors to influence both decisions and confidence in the same way. Against this assumption, asymmetries have been found in the influence that priors have on confidence compared to discrimination decisions. However, these patterns were found for high-level probabilistic expectations that are flexibly induced in the task context. It remains unclear whether this generalizes to low-level perceptual priors that are naturally formed through long term exposure. Here we investigated human participants' confidence in decisions made under the influence of a long-term perceptual prior: the slow-motion prior. Participants viewed moving line stimuli for which the slow-motion prior biases the perceived motion direction. On each trial, they made two consecutive motion direction decisions followed by a confidence decision. We contrasted two conditions – one in which the prior biased perceptual decisions and one in which decisions were driven by the sensory information alone. We found a confidence bias favoring the condition in which the prior influenced decisions, even when accounting for performance differences. Computational modeling revealed this effect to best be explained by confidence using the prior-congruent evidence as an additional cue, beyond the posterior evidence used in the perceptual decision. This suggests a confirmatory confidence bias favoring evidence congruent with low-level perceptual priors, revealing that, in line with high-level expectations, even long-term priors have a particularly strong influence at the metacognitive level.

Our perception of the environment is constantly compromised by uncertainties, and the Bayesian framework has become a popular way to account for how we cope with this uncertainty across a variety of cases (Kersten et al., 2004; Knill & Richards, 1996; Mamassian et al., 2002; McNamara et al., 2006; Yuille & Kersten, 2006; Zellner, 1995). According to Bayesian inference models of perception, our perceptual experience and decisions depend not only on incoming sensory information (the “likelihood”), but also on prior expectations (“priors”). These sources of evidence are combined in a precision-weighted way to form the “posterior” probability distribution function, on which the percept is based. More recently, these Bayesian decision models have also been extended to account for the sense of confidence that we have about the validity of our decisions, with confidence captured as the perceived posterior probability of being correct (Aitchison et al., 2015; Fleming & Daw, 2017; Kepecs & Mainen, 2012; Meyniel et al., 2015; Pouget et al., 2016; Sanders et al., 2016). This implies a role for priors in confidence as well, which has been supported empirically in both humans (Locke et al., 2020; Sherman et al., 2015, 2016) and non-human animals (Schmack et al., 2021).

Recent evidence suggests that priors may have dissociable impacts on perceptual decisions (first-order) and confidence judgments about those decisions (second-order). For example, priors about sensory precision, performance, or task difficulty have been shown to affect second-order without changing first-order performance (Olawole-Scott & Yon, 2023; M. Sherman & Seth, 2021; Van Marcke et al., 2024). Arguably, however, this does not constitute strong evidence for a differential impact of prior information on first- and second-order decisions: These higher-order priors concerned specifically the confidence-related features of the signal, but were irrelevant for perceptual decisions. Clearer evidence comes from recent work examining how decision-relevant priors are weighted in confidence relative to their use in discrimination decisions. This work found robust dissociations between these processing levels: Prior information was used to a greater extent in subjective confidence than in the decisions themselves (Constant et al., 2023). This suggests that priors have a particular role in confidence, over and above their effect on perceptual decisions. However, that work – along with most of the work on priors in confidence – manipulated priors based on probabilistic expectations that were flexibly induced in the task context and likely acted post-perceptually, which we will refer to as ‘high-level’. It is possible that these high-level priors especially influence confidence because abstract probabilistic information is difficult to integrate with perceptual information, or because it may even be advantageous not to rely on them for perceptual decisions when they flexibly and rapidly change.

In contrast to those high-level, flexible expectations, we also have low-level priors that act strongly at a perceptual level. These priors are often formed naturally across a breadth of life experience, such as a prior to perceive light as coming from above (Mamassian & Goutcher, 2001), and are sometimes therefore referred to as 'long-term' priors (e.g. Hardstone et al., 2021). Further, they have been suggested to act differently to high-level expectations in a variety of other ways, being slower to update, more cognitively impenetrable, and more context-independent (de Lange et al., 2018; Seriès & Seitz, 2013; Teufel & Fletcher, 2020), as well as potentially being implemented in a more bottom-up processing stream (Gerardin et al., 2010; Teufel & Fletcher, 2020) – although for counter-evidence, see (Hardstone et al., 2021). It therefore remains unclear whether the same dissociations between decisions and confidence under the influence of expectations would hold for long-term perceptual priors.

Here, we test this question by investigating confidence under the influence of a long-term naturally formed prior. To do so, we focus on the 'slow-motion' prior. Humans have a prior for motion speed to be slow, which is thought to originate from the fact that many objects we perceive in the world are stationary or move very little (Weiss et al., 2002). This has been shown to impact motion perception across different tasks and to explain a variety of biases in motion perception (Stocker & Simoncelli, 2004, 2007; Weiss et al., 2002; Zhang & Stocker, 2022). In line with Bayesian theory, this slow-motion prior influences inferences about motion particularly when sensory information is uncertain, or in other words when likelihoods are noisy. One example of uncertain motion information occurs when lines whose ends are poorly visible move through an aperture, creating what is called the "aperture effect". Under these conditions, the direction and speed of the line motion are inherently ambiguous and participants are initially biased to perceive a motion direction that is exactly orthogonal to the line orientation (Montagnini et al., 2007). This is consistent with perceiving the slowest possible motion that can still explain the sensory information, in line with the idea that the slow-motion prior influences their perceptual inference (Sotiropoulos et al., 2011; Weiss et al., 2002). Further supporting this, previous work found that this bias strongly influenced participants' motion direction decisions and found the influence to be well captured by a Bayesian decision model including the slow-motion prior (Sotiropoulos et al., 2011). They also found that this bias was reduced and even reversed by exposing participants to fast motion between test sessions, surprisingly revealing that even such low-level priors can be updated, in line with some other work (Adams et al., 2004). Here, we build on similar stimuli as Sotiropoulos and colleagues (2011) to investigate the effects of the slow-motion prior on decisions and confidence.

In a confidence forced-choice task, participants made two consecutive perceptual decisions followed by a confidence choice on every trial (Mamassian, 2020). Each perceptual decision consisted in reporting whether the perceived motion direction of a set of lines was clockwise or counterclockwise from a reference. We contrasted two conditions in order to examine the impact of information from the slow-motion prior on perceptual decisions and confidence judgments. One condition was tailored so that the slow-motion prior contributed strongly to whether the perceived motion direction was clockwise or counterclockwise to the reference, and the other condition was such that this was driven purely by the stimulus information. By matching the first-order decision behavior across these conditions, we could then examine whether there was a remaining confidence bias favoring one of them. If so, this could suggest that confidence uses the prior information differently than the first-order decisions do. Further, to quantify this effect and better understand the mechanisms underlying the use of perceptual priors in confidence, we fit and compared several models of confidence choices under the influence of priors.

Results

On every trial of the confidence forced-choice task, participants made two consecutive decisions about the motion direction of a matrix of lines moving through a circular aperture (Fig. 1A). The preferred motion direction due to the slow-motion prior was orthogonal to the orientation of the moving lines. On the outside of the circular aperture, two arcs, each subtending 90°, served as choice regions – one orange, one blue – separated at the 'reference' (Fig. 1A). In each interval participants made a decision about whether the motion direction was towards the displayed blue or orange area. Participants then made a confidence forced-choice on every trial about whether they were more confident in the first or second interval (Fig. 1B). We created two conditions. In the Bias condition, the preferred direction – orthogonal to the line orientation – fell within either the orange or blue area, creating a decision bias (Fig. 1A). In the No-Bias condition, the preferred motion direction(s) fell exactly 90° away from the reference and hence there was no bias favoring either decision. In this way, we could compare the No-Bias condition in which decision performance was determined entirely by the incoming sensory information (the likelihood) against the Bias condition, in which the slow-motion prior influenced performance. The strength of the likelihood, or the difficulty of the stimuli, was controlled by the angle, θ, subtended by the true motion direction and the reference.

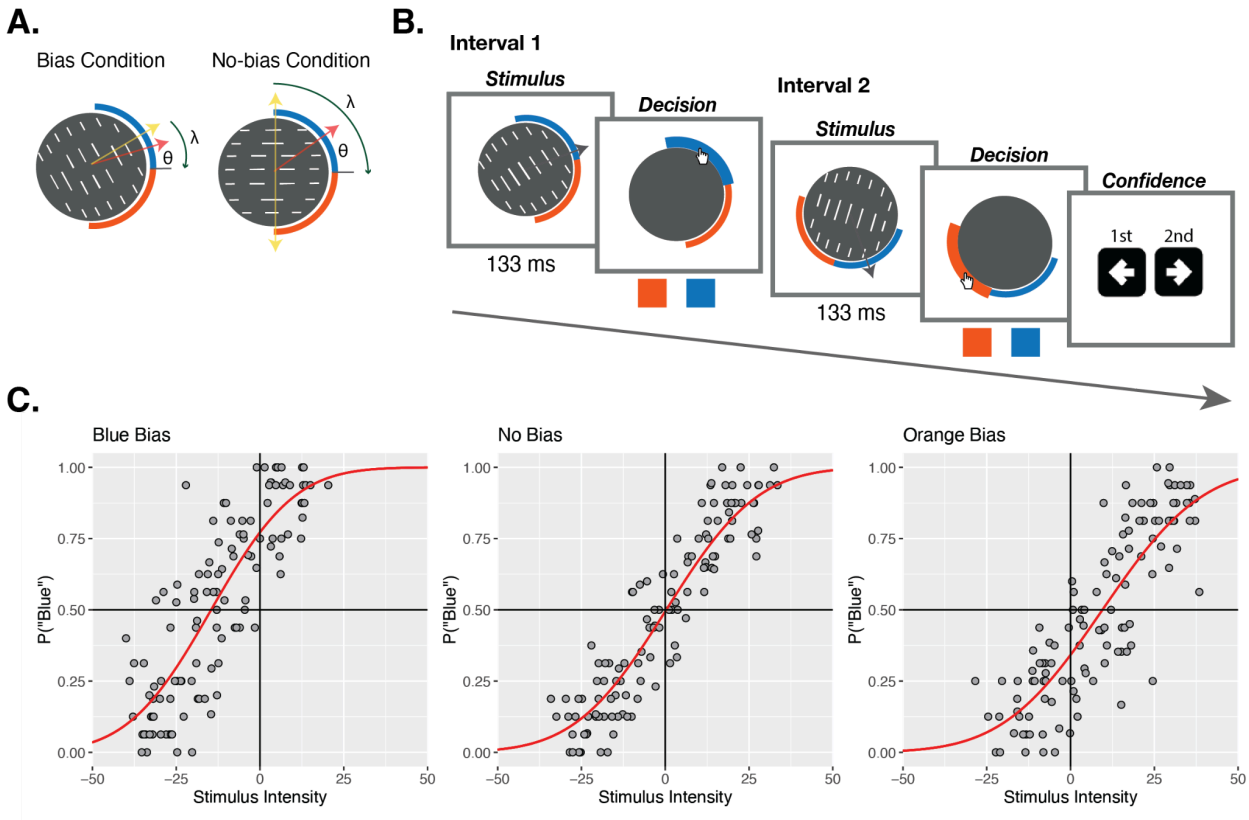

Figure 1. Task and Manipulations. A. Conditions and sketch of stimuli. The stimuli consisted of a matrix of lines moving through a circular aperture, shown at low contrast and for a short duration. The yellow arrows indicate the preferred motion directions in line with the prior, orthogonal to the line orientation. In the Bias condition, the angle, λ, between this prior and the reference was 35°, creating a decision bias favoring the color in which the prior falls. In the No-Bias condition, λ was 90°, such that we expected no prior-induced decision bias favoring either color. The red arrows indicate the true motion direction on an example stimulus. The strength of the likelihood was controlled by the angle, θ, between this motion direction and the reference. A larger θ was required in the No-Bias condition to reach the same choice rates, since there was no contribution of the prior to the orange/blue decision. **B. Experimental task.** A trial of the confidence forced-choice task included two intervals, in which participants saw a stimulus and made an orange/blue decision about the motion direction of the lines, followed by a confidence forced-choice about which of their responses they were more confident about. **C. Manipulation check.** Pooled results from the adaptive staircase (ASA) procedure in which many different θ values were tested in their effect on first-order decisions in both conditions before the full confidence task. Each gray point corresponds to the responses for a given stimulus intensity level from one participant. θ values towards the blue region are encoded as positive, and towards the orange region are encoded as negative. Red psychometric functions show the fit cumulative normal distribution function capturing the relationship between θ and the probability of choosing blue. In the No-Bias condition (middle panel), there is no choice bias for orange or blue. In the Bias condition, when the prior falls in the blue region (left panel), the function is shifted to the left, showing a successfully induced choice bias favoring blue. When the prior falls in the orange region (right panel), the function is shifted to the right, showing a successfully induced choice bias favoring orange.

Manipulation Check

We predicted that the slow-motion prior would bias participants' decisions in the Bias condition, and hence that the Bias condition would require a smaller θ in order to lead to the same decision rates as the No-Bias condition. This effect would serve as a basic manipulation check to ensure that the conditions and slow-motion prior had the intended impact. Additionally, it was necessary to quantify this effect in each participant before the experiment such that we could choose the θ values that would be expected to lead to matched decision rates. To do this, before the main task we ran a staircasing procedure which sampled different θ values in both conditions and then fit psychometric functions to this data. The fit psychometric functions capturing the bias effect across all pooled participants are shown in Fig. 1C. Motion into the orange area was encoded with negative θ's and blue area with positive θ's. We found a clear bias effect, such that the psychometric was shifted leftward when the preferred motion direction from the prior was in the blue region and rightward when it was in the orange region, compared to the No-Bias condition (Fig. 1C). The point of subjective equality (PSE) when there was a blue bias was -14.57°, meaning that when the line motion direction was 14.57° into the orange region, participants still chose orange and blue at equal rates. In contrast, the PSE when there was an orange bias was 9.53°, and the PSE in the No-Bias condition was 0.30°. Together, this suggests that the bias manipulation worked as planned.

Motion-Direction Decisions

After running the staircasing procedure, we chose the values of θ that corresponded to a 0.65 and 0.85 probability of choosing each color for each participant. This procedure was followed both for the No-Bias condition and for the Bias condition (for a total of eight θ values). In the Bias condition, we chose values corresponding to a 0.65 and 0.85 probability of choosing blue when the slow-motion prior induced a blue bias, and a 0.65 and 0.85 probability of choosing orange when it induced an orange bias. Overall, this meant that we chose stimulus levels that were expected to lead to matched choice probabilities between conditions. Crucially, however, the choice rates in the Bias condition would be in part due to the slow-motion prior; whereas the choice rates in the No-Bias condition would be solely driven by the stimulus evidence. These four θ values in each condition were paired in all possible combinations for the two intervals of the confidence forced-choice task, except for pairs repeating the identical θ in the same condition.

To confirm that the choice rates were matched overall across the two conditions, we ran a logistic mixed effects model on Response including Expected P('Blue') (as an index of the stimulus intensity setting), Condition (Bias vs No-Bias), and their interaction as fixed effects, as

well as by-participant random intercepts. If the chosen θ values from the staircasing procedure failed to produce matched choice rates between conditions, this will generate an interaction between Expected P('Blue') and Condition. We found this significant interaction, $\chi^2(3)=302.97$, $p<0.001$, revealing that the response rate per stimulus setting additionally depended on the condition. The response rates were more extreme in the Bias condition (Fig. 2A), with higher odds of choosing blue when the stimulus was blue (Expected P('Blue') of 0.65: $OR_{Bias/No\text{-}Bias}$ = 1.55, 95% CI [1.32, 1.81], Z=8.75, p<0.001; Expected P('Blue') of 0.85: $OR_{Bias/No\text{-}Bias}$ = 1.13, 95% CI [0.92, 1.38], Z=1.81, p=0.071), and lower odds when the stimulus was orange (Expected P('Blue') of 0.15: $OR_{Bias/No\text{-}Bias}$ = 0.54, 95% CI [0.42, 0.69], Z=-7.91, p<0.001; Expected P('Blue') of 0.35: $OR_{Bias/No\text{-}Bias}$ = 0.49, 95% CI [0.41, 0.58], Z=-13.43, p<0.001), compared to the No-Bias condition, although this contrast was not significant for the strongest blue setting. The more extreme response rates in the Bias condition suggests that the bias effect got stronger compared to what was seen in the staircasing procedure. To quantify this, we estimated the strength of the bias effect in the experiment by fitting a cumulative normal distribution function with one symmetrical bias level (+/- μ_{Bias}) for the orange and blue directions relative to the No-Bias condition, which was fit separately as a baseline (Fig. 2A). The sensory sensitivity was also allowed to differ between conditions. This was done for each participant and the resulting mean μ_{Bias} across participants was 20.81° (Fig. 2A), indeed indicating a stronger bias effect than in the staircasing (Fig. 1C). The resulting difference in first-order response rates must be accounted for when testing for a confidence bias.

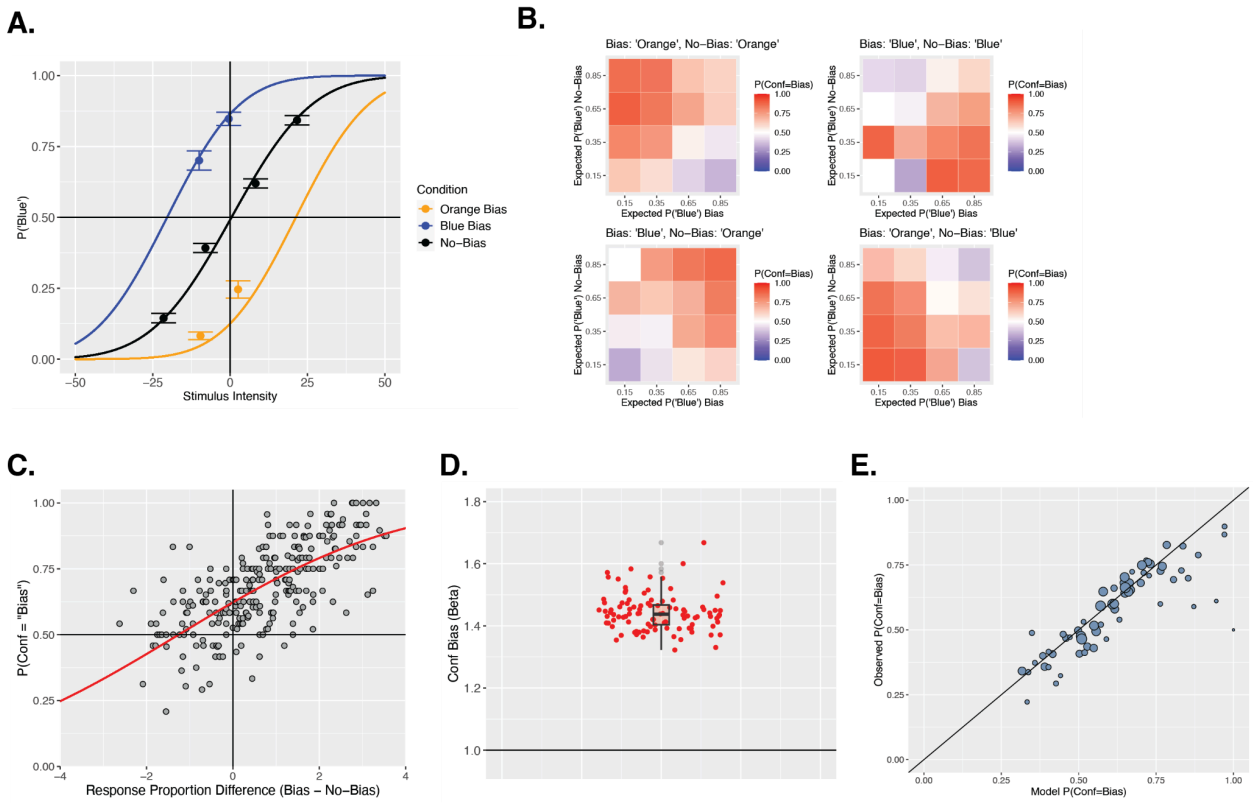

Figure 2. Motion-Direction and Confidence Decision Results. A. Motion-direction decisions. Choice probability results for the first-order orange/blue decisions in each condition. The four intensity levels were targeting 0.15, 0.35, 0.65 and 0.85 probabilities of choosing blue in each condition, although the resulting choice rates were less extreme in the No-Bias condition than in the Bias condition. Error bars reflect the SEM across participants. The psychometric functions were fit to each individual participant to capture the strength of the bias from the slow-motion prior relative to the No-Bias baseline, and to capture possible differences in sensitivity between conditions. The functions shown here reflect the mean fit bias (μ) and noise (σ) values across participants. The mean μ_{Bias} was 20.81, mean $\mu_{No\text{-}Bias}$ was 0.44, mean σ_{Bias} was 18.50, and mean $\sigma_{No\text{-}Bias}$ was 20.46. **B. Confidence forced-choices.** Confidence forced-choice results for each pair of stimulus settings – captured by the expected choice rates, as this is how they were chosen from the ASA procedure – with one interval in each condition, grouped by possible combination of orange/blue responses. The color indicates the probability of a confidence choice favoring the Bias condition interval. The red dominance overall suggests that participants are more often most confident in the Bias condition interval. **C. Non-parametric confidence bias results.** On the x axis is the difference in the rate of choosing the expected color between the Bias and No-Bias condition, z-scored. Each gray point represents this difference for one pair of stimulus settings with one interval in each condition, from one participant. On the y axis is the probability of choosing the Bias condition interval as the more confident interval. When there is no difference in response proportions between conditions (x=0), and the choice rates are therefore matched for that pair of settings, we expect equal confidence choices between conditions. Otherwise, this indicates a confidence bias. The red psychometric function captures the fit cumulative normal distribution function to the relationship between this response proportion difference and the confidence choice rates. The PSE is shifted leftward to -1.26, indicating a confidence bias favoring the Bias condition. **D. CFC-model confidence bias results.** We fit the *cfc-model* across 100 bootstrapped runs and extracted the fit confidence bias parameter β from each. Each red point corresponds to one fit β parameter result. The boxplot shows the median, interquartile range (IQR) with

hinges showing the first and third quartiles, and vertical whiskers stretching to most extreme data point within 1.5*IQR from the hinges. The horizontal black line at y=1 shows the expected β if there was no confidence bias. **E. CFC-model fit.** Predicted against observed confidence choice rates across all participants. Each point reflects a pair of stimulus settings with one interval in each condition and a given set of two orange/blue responses, and the size of the point reflects the number of trials. The closer the points are to the x=y line, the better the model is at predicting confidence choice rates.

Confidence Forced-Choices

We then sought to explore whether, after accounting for any potential difference in orange/blue decision rates between conditions, there would still be a confidence bias favoring one of them. If so, this might indicate a different influence of the prior in confidence, which could then be further explored using Bayesian confidence modeling. The raw confidence forced-choice rates are shown per stimulus setting and pair of possible responses in Fig. 2B, and suggest a higher rate of confidence choices favoring the Bias condition overall. Because the orange/blue decision rates were not matched across conditions, and following our pre-registered plan for this potential situation, we used a non-parametric approach to assess confidence bias, which accounted for potential mismatch in first-order decision rates. We took the difference in response rates for the preferred stimulus between conditions for every stimulus setting pair, excluding pairs with both intervals in the same condition, and z-scored them (see Toscani et al., 2021). For example, for the stimulus settings expected to lead to a 0.85 probability of choosing blue in each condition, we took the normalized difference between the observed rates of choosing blue between the conditions. A normalized difference of zero then indicated matched first-order choice rates, positive differences indicated more extreme choice rates in the Bias condition, and negative differences indicated more extreme choice rates in the No-Bias condition. Then, regardless of whether these values were shifted positively overall, as we expect due to the result of more extreme choice rates in the Bias condition, we can fit a cumulative normal distribution function to the relationship between these normalized response proportion differences and the probability of choosing the Bias condition as the more confident interval. If the PSE is at 0, this indicates that when participants have matched first-order choice rates they also have matched confidence choice rates between conditions, suggesting no confidence bias. If, however, the PSE is shifted away from 0, this indicates a confidence bias. We found a clear negative shift of this function, fit to the pooled participants, with a PSE of -1.26, suggesting a confidence bias favoring the Bias condition (Fig. 2C). When response rates are matched (normalized response proportion difference = 0), the fit psychometric predicts the participants to still choose the Bias condition as the more confident interval at a rate of 0.62.

Quantifying Confidence Bias: CFC-Model

To further quantify this confidence bias, we fit the *cfc-model* developed in previous work (Mamassian & de Gardelle, 2022) for use with the confidence forced-choice paradigm. This model includes a confidence bias term, β, which, for a single task or condition, scales the estimated sensory sensitivity of a participant, such that β=1 captures correctly estimated sensitivity, β>1 captures an overestimation of sensitivity – indicating overconfidence –, and β<1 captures an underestimation of sensitivity – indicating underconfidence. With two different conditions, the model fits β as the ratio of biases between them (one can only estimate whether participants are over- or under-confident in one condition relative to the other). In our case, we choose the 'No-Bias' condition as baseline, so a fit β larger than 1 would indicate an overconfidence in the 'Bias' condition. We fit the model to the pooled normalized data from all participants across 100 bootstrapped runs, which revealed the mean fit β to be 1.44 (SEM=0.006; Fig. 2D). The model fit is shown in Fig. 2E. These results, in line with the non-parametric analysis above, indicate a confidence bias favoring the Bias condition. While this model can be used to naively quantify a confidence bias relative to first-order choice rates, it does not provide an account of the computations occurring in our design, under the influence of the slow-motion prior. To investigate these computations, we consider a Bayesian decision and confidence model, along with several alternative generative models.

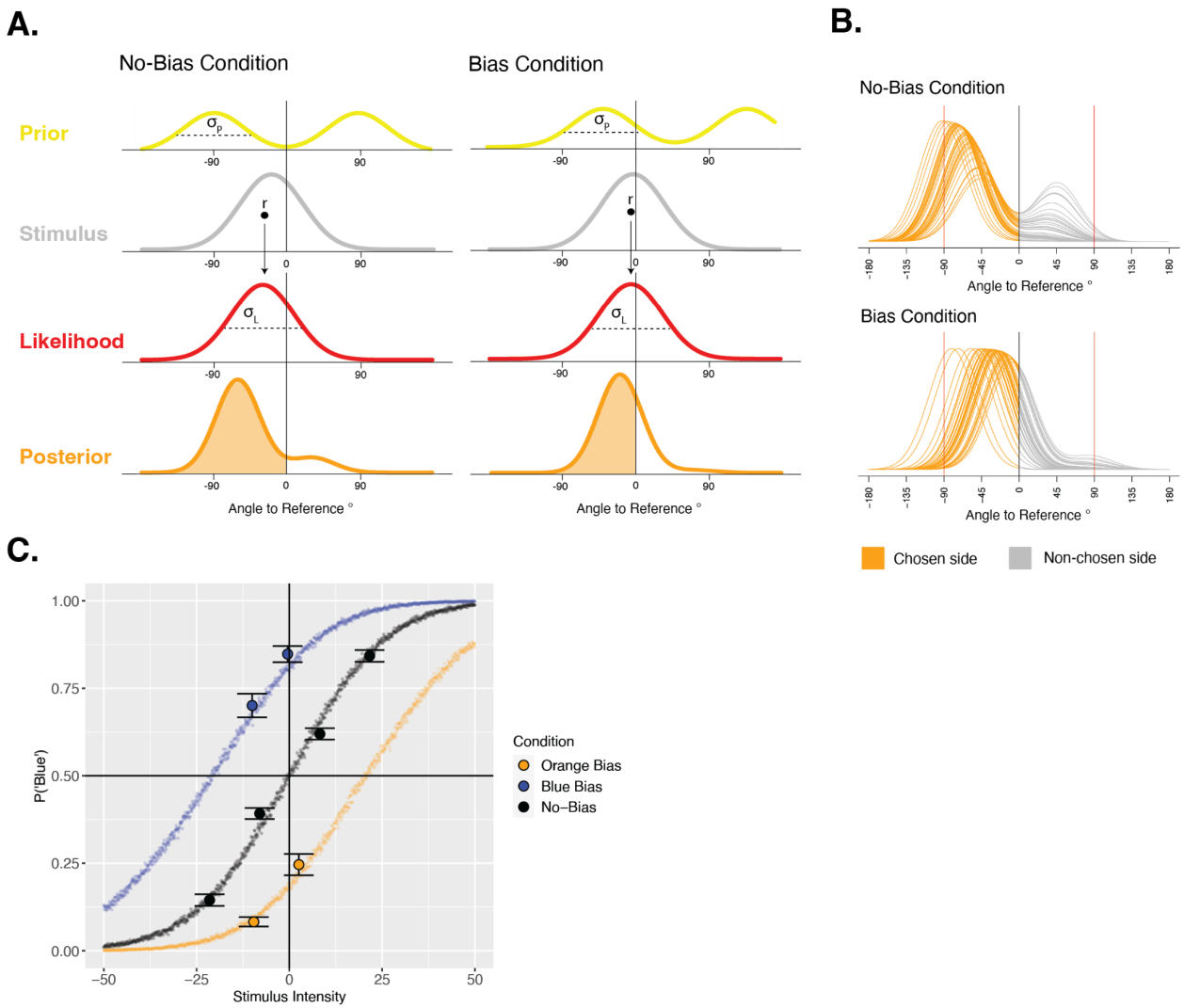

Figure 3. Bayesian Optimal Observer Model. A. Sketch of Bayesian Optimal Observer model. In the Bayesian model, the prior on perceived motion direction is a sum of two Gaussians of equal standard deviation (σ_P), each with a mean at one of the preferred, orthogonal motion directions. In the No-Bias condition these means are 90° away from the reference in either direction and in the Bias condition one is 35° towards the bias direction, and the other is 145° in the opposite direction, outside the allowed choice region (from -90° to 90°). The likelihood comes from the internal signal, *r*, sampled on a given trial from the stimulus with added internal noise (σ_L). The prior and likelihood are then combined to form the posterior distribution. The shaded orange area under the posterior then leads to the first-order decision and confidence. **B. Posteriors per condition from Bayesian Optimal Observer model.** Resulting posterior distributions from the Bayesian Optimal Observer model across 50 different samples of an internal signal from the stimulus to form the likelihood, only including samples that led to an orange (negative) decision. These reflect the posteriors across different trials of the same stimulus intensity and condition, which led to the same response. Red vertical lines indicate the allowed orange/blue choice region. The posteriors are colored orange on the chosen side, and gray on the non-chosen side. The confidence in these trials is captured by the ratio of area under the orange part of the posterior to the area under the gray part of the posterior, between the red vertical lines. Because the prior tends to pull the maximum of the posterior in the No-Bias condition away from the reference and towards -90°, the ratio of area under the orange versus gray parts of those posteriors tends to be higher, suggesting higher

average confidence in the No-Bias condition despite matched choice rates. **C. Modeled motion-direction decision results.** Results from the fit Bayesian Optimal Observer decision model, simulated across 1001 stimulus intensities ranging continuously (in steps of 0.1) from -50 to 50 using the best-fitting σ_P and σ_L parameters for each participant, against the data.

Bayesian Optimal Observer Model

In the Bayesian Optimal Observer model, we defined the prior to be the sum of two Gaussians of equal variance, each centered around one of the preferred orthogonal motion directions (Fig. 3A). In the No-Bias condition, these means were 90° away from the reference in either direction. In the Bias condition, one mean was 35° towards the bias direction, and the other was 145° away from the reference in the opposite direction, and was hence outside of the allowed decision region. The likelihood was a single Gaussian centered around the internal signal generated from the stimulus with added internal noise. These combined to form the posterior distribution (Fig. 3A). Orange/blue decisions in this model are based on the area under the posterior on either side of the reference within the allowed decision region (from -90° to +90°). When this area is larger on the negative side, an “orange” choice is made, and when it is larger on the positive side, a “blue” choice is made (Fig. 3A-B). Following the literature capturing Bayesian confidence as the perceived posterior probability of being correct (Aitchison et al., 2015; Fleming & Daw, 2017; Kepecs & Mainen, 2012; Meyniel et al., 2015; Pouget et al., 2016; Sanders et al., 2016), confidence in the decision is then proportional to that area, and the confidence forced-choice is based on whichever of these confidence values (the area under the posterior on the chosen side) is larger. The free parameters of this Bayesian model are the standard deviation of each of the two Gaussians that form the Gaussian mixture prior, σ_P, and the standard deviation of the likelihood, σ_L, that represents the sensory uncertainty. These were fit individually for each participant to find the parameters that could best explain their orange/blue decisions. We used a simplification to deal with computational complexity and considered the prior in the Bias condition to be a single Gaussian centered around 35° towards the bias direction. Since the other mean fell well outside the allowed decision region, this had a negligible impact. The results from these model fits are shown in Fig. 3C by simulating from the full unsimplified model, showing that the fit σ_P and σ_L parameters can still well account for motion-direction decisions.

We then used this fit model and investigated the confidence forced-choice patterns that would be expected given the first-order choice rates that we found. It is important to note that in this model, because of the shape of the Gaussian mixture priors, the distribution of posteriors across different samples of the stimulus (or in other words across trials) is quite different between the conditions (Fig. 3B). In the No-Bias condition, the maxima of these posteriors are

pulled towards the peaks of the prior at either + or -90°, so a lot of the area is distributed relatively far from the reference on the chosen side (Fig. 3B). Having a higher proportion of the area on the chosen side like this leads to higher confidence, and therefore the Bayesian optimal observer will on average have higher confidence in the No-Bias condition despite matched first-order choice rates. This pattern is shown clearly in the model simulations of the optimal Bayesian confidence observer given the first-order decision results (Fig. 4A). This suggests that, if our participants used the information from the slow-motion prior in a Bayesian optimal manner for their confidence, there would have been a confidence bias favoring the No-Bias condition, in the opposite direction from what is seen in the data (Fig. 4B). So, we next explore some alternative confidence models that might better explain the patterns found.

Distance-to-Criterion Confidence Model

In the Bayesian Optimal confidence model, confidence is based on the area under the full Bayesian posterior. However, another possibility for computing confidence is to base it off of a single sample from the posterior, in which case the brain may not need knowledge of the full distributional properties (see, e.g., (Locke et al., 2022)). In this model, confidence is proportional to the distance between the posterior sample and the reference, or ‘criterion’ (Fig. 4C). Samples further from the criterion very clearly support that choice and lead to higher confidence and samples close to the criterion indicate uncertain choices and lead to lower confidence. This confidence model has no free parameters beyond the first-order sensitivity parameters fit to perceptual decisions and used in all models. We simulated confidence forced-choices from it to investigate whether it could better account for the data. This revealed even more extreme predictions of a confidence bias favoring the No-Bias condition than in the Bayesian Optimal model (Fig. 4D), suggesting this model could only poorly capture the confidence data.

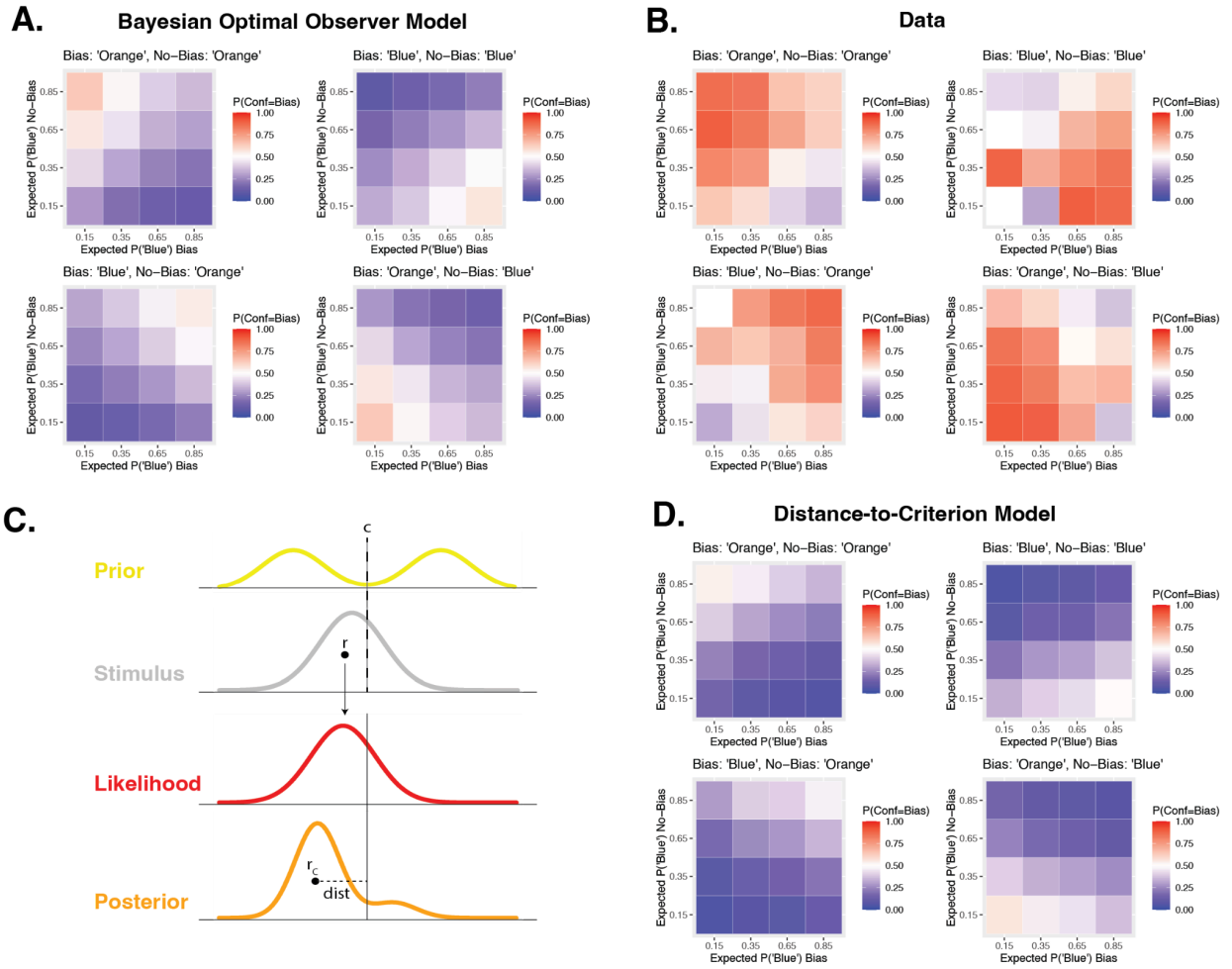

Figure 4. Bayesian Optimal and Distance-to-Criterion Confidence Models. A. Confidence predictions from Bayesian Optimal Observer model. Confidence forced-choice predictions simulated from the Bayesian Optimal Observer model for each pair of stimulus settings – captured by the expected choice rates – with one interval in each condition, grouped by possible combination of orange/blue responses. The color indicates the probability of a confidence choice favoring the Bias condition interval. The blue dominance overall suggests that this model predicts more confidence choices favoring the No-Bias condition, unlike what is seen in the data. **B. Confidence forced-choice data.** Observed confidence forced-choice results for each pair of stimulus settings with one interval in each condition, grouped by possible combination of orange/blue responses. **C. Sketch of Distance-to-Criterion model.** In the Distance-to-Criterion model, the prior, likelihood, and resulting posteriors are the same as in the Bayesian Optimal Observer model. Motion-direction decisions also work identically. However, at the confidence level, instead of computing the full Bayesian posterior probability of being correct, this model bases confidence on the absolute distance, *dist*, between a single sample from the posterior, r_c, and the reference, or criterion (c). Larger distances lead to higher confidence and smaller distances lead to lower confidence. **D. Confidence predictions from Distance-to-Criterion model.** Confidence forced-choice predictions simulated from the Distance-to-Criterion model for each pair of stimulus settings with one interval in each condition, grouped by possible combination of orange/blue responses. Like in the Bayesian Optimal Observer model, the blue dominance overall suggests that this model predicts more confidence choices favoring the No-Bias condition, unlike what is seen in the data.

Overweighted Prior Confidence Model

The Bayesian Optimal and Distance-to-Criterion models both base the confidence on the same posterior as the first-order decisions, and fail to capture the confidence bias seen in the data favoring the Bias condition. It is possible that the confidence data can be better explained by using a different weighting of the prior relative to likelihood in computing confidence, similar to the pattern found in previous work with priors being more strongly weighted in confidence than in first-order decisions (Constant et al., 2023). In the Overweighted Prior (OWP) model, the variance of the likelihood is scaled by a multiplicative factor w during the confidence computation to capture a relative over- or underestimation of the likelihood precision and hence an over- or underweighting of the likelihood information relative to the prior (Fig. 5A). When $w=1$, the likelihood variance is correctly estimated, the relative weighting is optimal, and this model is identical to the Bayesian Optimal confidence model. When $w>1$, the likelihood variance is overestimated and therefore it has a weaker effect and the prior is relatively overweighted in confidence. When $w<1$, the likelihood variance is underestimated and therefore it has a stronger effect and the prior is relatively underweighted in confidence. We fit this model with its free parameter w to the group data using a maximum likelihood estimation approach. This revealed a best-fitting weighting parameter of $w=1.54$, indicating that the prior is overweighted in confidence, in line with previous work using a similar model implementation (Constant et al., 2023).

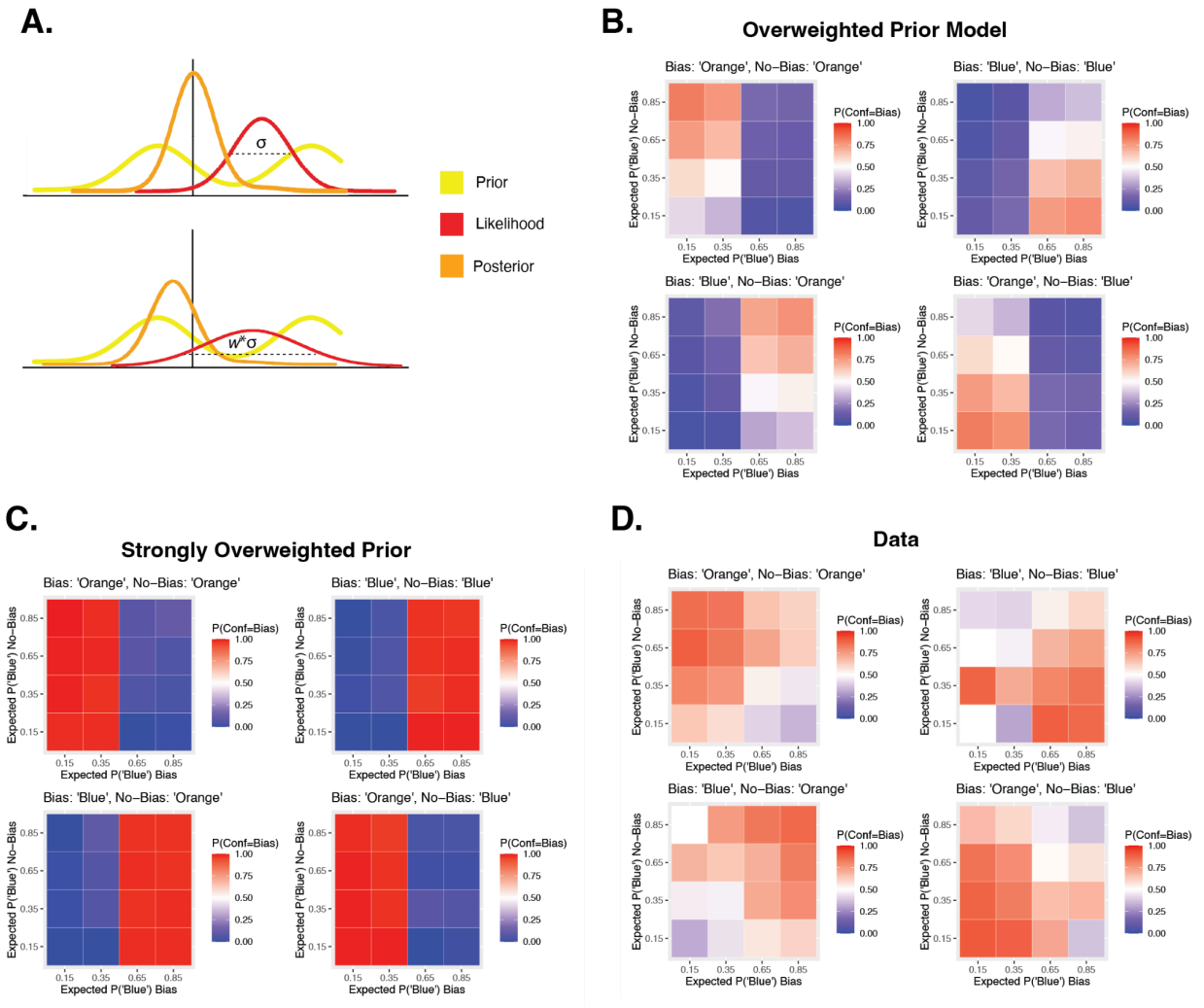

Figure 5. Overweighted Prior Confidence Model. A. Sketch of Overweighted Prior model. In the Overweighted Prior model, confidence is computed the same way as in the Bayesian Optimal Observer model, by taking the area under the posterior distribution on the chosen side. However, in forming this posterior, participants over- or underestimate the precision of the likelihood relative to the prior, leading to the prior being relatively over- or underweighted in the confidence computation. The over- or underestimation of the likelihood precision is captured by a scaling term *w* that acts as a multiplicative factor on the variance of the likelihood. When this value is 1, the variance is correctly estimated and this model is the same as the Bayesian Optimal Observer model. When *w*<1, the prior is relatively underweighted and when *w*>1 the prior is relatively overweighted in confidence. **B. Confidence predictions from fit Overweighted Prior model.** Confidence forced-choice predictions simulated from the fit Overweighted Prior model, with the best-fitting *w*=1.54, for each pair of stimulus settings – captured by the expected choice rates – with one interval in each condition, grouped by possible combination of orange/blue responses. The color indicates the probability of a confidence choice favoring the Bias condition interval. More red in these simulations compared to the Bayesian Optimal Observer model (Fig. 4A) suggests that overweighting prior information can lead to more confidence decisions favoring the Bias condition, closer to what is seen in the data. **C. Confidence predictions of extreme prior overweighting.** Confidence forced-choice predictions simulated from the Overweighted Prior model using an extreme overweighting, *w*=4, for each pair of stimulus settings with one interval in each condition, grouped by possible combination of orange/blue responses. As the overweighting gets extreme, this

model predicts the Bias condition to be the confident interval whenever the orange/blue response is in line with the prior, and the No-Bias condition to be the confident interval whenever the orange/blue response goes against the prior. **D. Confidence forced-choice data.** Observed confidence forced-choice results for each pair of stimulus settings with one interval in each condition, grouped by possible combination of orange/blue responses.

Simulations from this fit model are shown in Fig. 5B. These demonstrate the ability of this model to better capture the confidence choices favoring the Bias condition than the Bayesian Optimal and Distance-to-Criterion models. However, this model predicts confidence to be particularly high in the Bias condition when the first-order response is in line with the prior, and to be particularly low when the first-order response goes against the prior. This leads to the prediction that confidence choices favoring the Bias condition will occur when the Bias condition response was in line with the bias direction, and confidence choices favoring the No-Bias condition will occur when the Bias condition response went against the bias direction. As w gets higher and the prior is weighted more strongly, this pattern gets more extreme, with confidence choices almost always favoring the Bias condition following responses in line with the prior, and almost always favoring the No-Bias condition otherwise (Fig. 5C). This model still cannot qualitatively capture the confidence choice patterns seen in the data (Fig. 5D), which do not follow this split.

Prior-Congruent Evidence (PCE) Confidence Model

From the previous models, it seems that confidence forced-choices are biased towards the condition that is more driven by the slow-motion prior, but not simply by overweighting the prior in the integration to form the posterior. Another possibility is that the information from the prior forms its own cue to confidence, separately from the Bayesian posterior evidence used in the first-order decisions. The Prior-Congruent Evidence confidence model suggests that confidence is based solely on the amount of prior-congruent evidence (PCE) available in the line motion stimulus, independently of the posterior evidence or the orange/blue decision made. In the stimuli used here, the prior-congruent evidence is the component of the motion information that goes in the preferred orthogonal motion direction. So, confidence here was proportional to the length of this vector component, shown in Fig. 6A, with no additional free parameters for confidence in this model. There tended to be more of this prior-congruent evidence in the Bias condition, since line motion directions were closer to the preferred orthogonal direction, and hence simulations from this model reveal a confidence bias favoring the Bias condition (Fig. 6B), in line with the confidence forced-choice results. However, the predicted confidence bias from this model is stronger than what we see in the data.

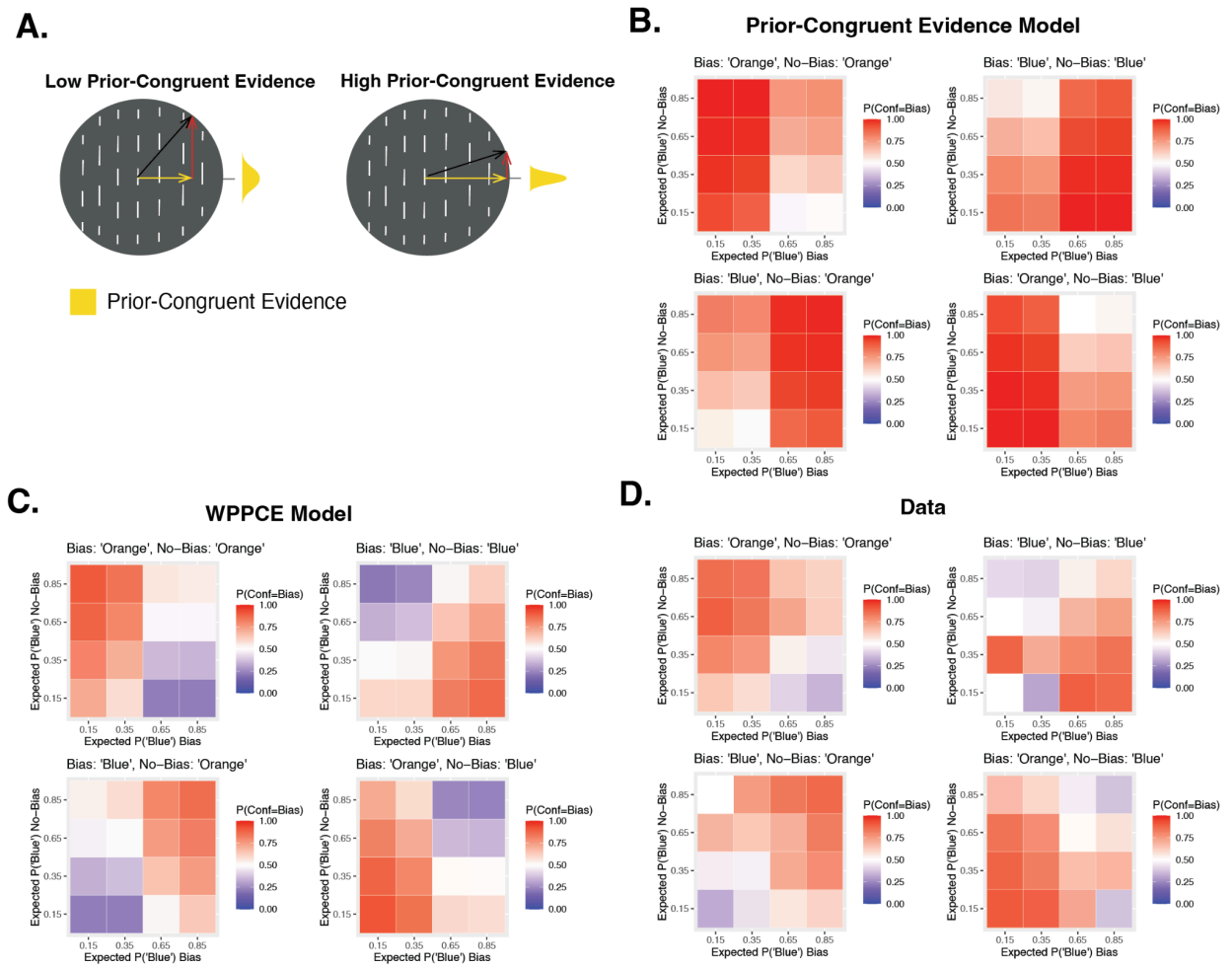

Figure 6. Prior-Congruent Evidence Models. A. Sketch of prior-congruent evidence. In the line-motion stimuli, the evidence that is congruent with the prior is the component of the motion that is in the preferred, orthogonal motion direction, shown as the yellow vector component. Stimuli with high PCE have a motion direction that is closer to the preferred one, and therefore a longer vector component along that dimension. In the PCE confidence model, confidence was directly proportional to this PCE, and was not based on the posterior evidence or on the orange/blue decision made. **B. Confidence predictions from Prior-Congruent Evidence (PCE) model.** Confidence forced-choice predictions simulated from the PCE model for each pair of stimulus settings – captured by the expected choice rates – with one interval in each condition, grouped by possible combination of orange/blue responses. The color indicates the probability of a confidence choice favoring the Bias condition interval. The red dominance overall suggests that this model predicts more confidence choices favoring the Bias condition, in line with, though more extreme than, what is seen in the data. **C. Confidence predictions from Weighted Posterior and Prior-Congruent Evidence (WPPCE) model.** Confidence forced-choice predictions simulated from the WPPCE model with best-fitting α=0.34 for each pair of stimulus settings with one interval in each condition, grouped by possible combination of orange/blue responses. In the WPPCE model, confidence is based on a weighted combination of the PCE, as shown in (A), as well as the Bayesian Optimal posterior evidence, as shown in Fig. 3A. **D. Confidence forced-choice data.** Observed confidence forced-choice results for each pair of stimulus settings with one interval in each condition, grouped by possible combination of orange/blue responses.

Weighted Posterior and Prior-Congruent Evidence (WPPCE) Confidence Model

The PCE model reveals that basing confidence on the prior-congruent evidence in the stimulus leads to a confidence bias favoring the Bias condition. However, when confidence is based solely on that PCE, the bias effect is stronger than what we see in the data. Next we consider a model in which this PCE acts as an additional confidence cue, but not the sole information on which confidence choices are based. In the WPPCE model, confidence is based on a weighted combination of the posterior evidence, which the first-order decisions are based on, as well as the degree of prior-congruent evidence. The influence of each of these two cues is determined by the weighting parameter, α, such that confidence is proportional to the sum of (1-α)*Posterior Evidence and α*PCE. Higher α values then lead to more contribution from the PCE. When α=1, the PCE is the only cue to confidence and this is identical to the PCE model, and when α=0, the Bayesian posterior evidence is the only cue to confidence and this is identical to the Bayesian Optimal Observer model. We fit this free weighting parameter to the group data using a maximum likelihood estimation approach, and found that a value of α=0.34 could best explain the data. Because the posterior evidence and PCE use different units, it is difficult to interpret the raw α value. However, simulations from this fit model are shown in Fig. 6C, and can capture the data well qualitatively, showing a similar confidence bias for the Bias condition that is not solely split based on whether the response was prior-congruent (Fig. 6C-D).

Model Comparison

To compare the models quantitatively, we computed the AIC for each, therefore also penalizing the models with more free parameters. This revealed an $AIC_{\text{Bayesian Optimal}}$ = 30763.85, $AIC_{\text{Distance-to-Criterion}}$ = 38770.61, AIC_{OWP} = 29838.48, AIC_{PCE} = 31676.98, and AIC_{WPPCE} = 25851.11 (Fig. 7A). The best model to explain the group confidence forced-choice data is the WPPCE confidence model (Fig. 7B), in which confidence choices are based not only on the posterior evidence that formed the first-order decision, but additionally on the prior-congruent evidence as a separate cue. We additionally fit the five models to each individual participant and compared them. This revealed the WPPCE confidence model to be the winning model in 23 out of the 24 participants, with one participant for whom the PCE model had the lowest AIC, although this was inconclusive in comparison to the WPPCE (Fig. 7C). Taken together, these results suggest that, relative to their first-order behavior, participants have a confidence bias that favors the condition in which the slow-motion prior more strongly influences decisions. Further, this bias can best be explained by a model in which they use the prior-congruent information as an

additional cue to inform their confidence, in combination with the Bayesian posterior evidence used for the motion-direction decisions.

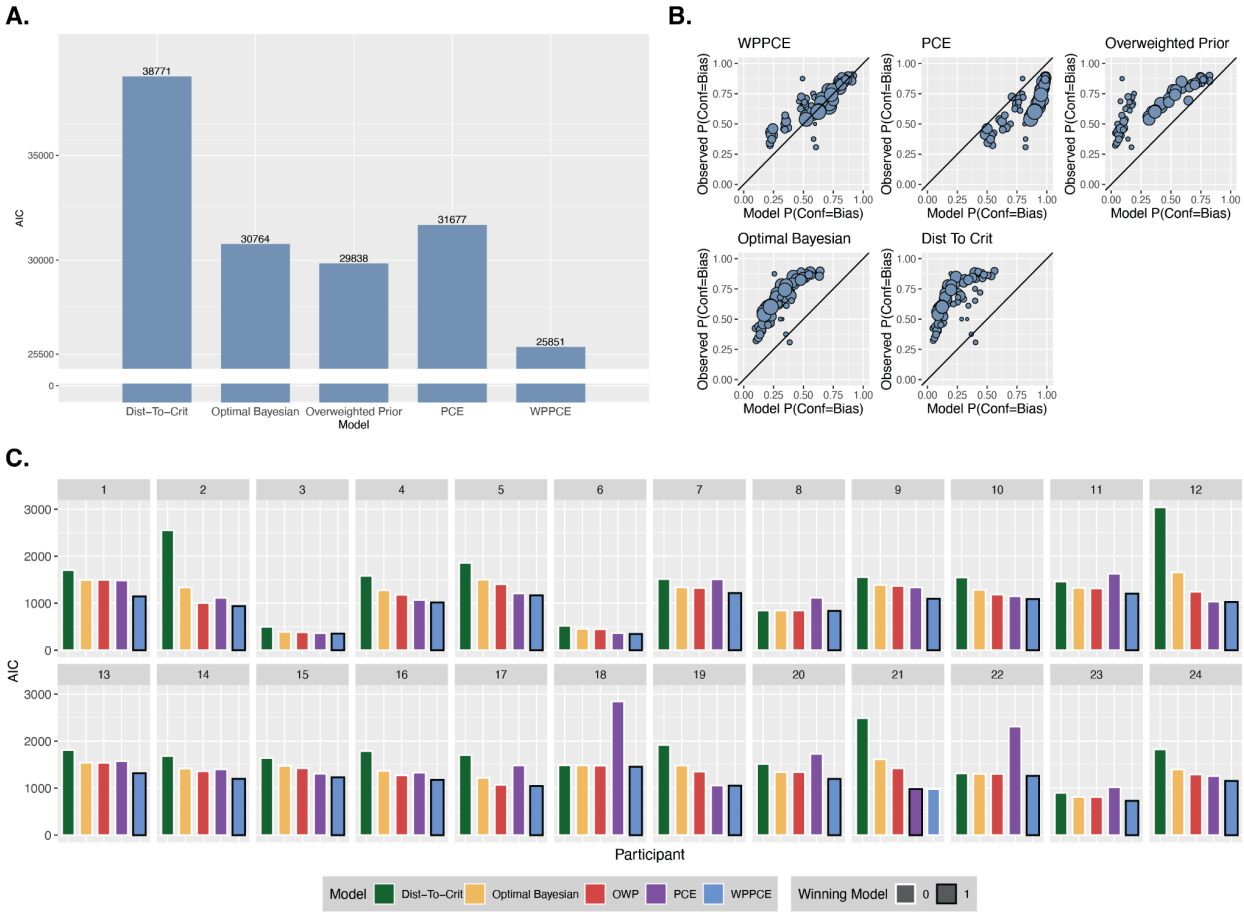

Figure 7. Model Comparison. A. AIC group results. AIC comparison of models fit to the pooled group data. **B. Model goodness of fit.** Predicted against observed confidence choice rates from each of the models. Each point reflects a pair of stimulus settings with one interval in each condition and a given set of orange/blue responses, and the size of the point reflects the number of trials. The closer the points are to the x=y line, the better the model is at predicting confidence choice rates. Confidence choice rates reflect the probability of choosing the Bias condition as the more confident interval. So, points above the identity line reflect the model underpredicting confidence choices for the Bias condition, and points below the identity line reflect the model overpredicting confidence choices for the Bias condition. **C. Subjectwise AIC results.** AIC comparison of models fit to each individual participant. The winning model for each participant is shown outlined in black. This was the WPPCE model for all except one participant.

Discussion

Previous work has revealed dissociations between the impact that priors have on confidence versus first-order decisions, suggesting a particularly strong role of priors in shaping confidence. However, this has previously only been examined in high-level priors that are

induced flexibly in a task context, and may not be generalizable to the perceptual priors that are formed naturally across the lifetime. Here, we investigated the role of priors in perceptual decisions and confidence using a long-term perceptual prior, namely, the slow-motion prior. Using moving line stimuli we created two conditions: one in which motion direction decision performance was strongly biased by the slow-motion prior (Bias condition), and another in which no bias was induced and performance was driven by the stimulus information (No-Bias condition). We found that, even when any differences in first-order decision rates between conditions were accounted for, there was a confidence bias towards the Bias condition. This suggests confidence to favor the condition with more influence from the prior on decisions, rather than the condition with a more informative stimulus. By comparing computational models, we revealed this effect to best be explained by confidence choices using the degree of prior-congruent evidence as an additional cue, over and above the posterior evidence that is used in first-order decisions.

In line with findings using high-level, flexibly induced priors (Constant et al., 2023), the results here reveal that even long-term perceptual priors have a different influence on first-order performance versus confidence, with prior-congruent information having a particularly strong impact at the metacognitive level. This underlines a striking consistency of this confidence-performance dissociation across different types of priors. Because these priors are often argued to act and even be implemented differently (de Lange et al., 2018; Teufel & Fletcher, 2020), finding this dissociation to hold in the case of long-term priors is enlightening about its nature. Based on work using instructed probabilistic priors, it remained unclear whether their disproportionate impact on confidence was due to the abstract probabilistic information being very unlike (and therefore difficult to integrate with) perceptual information, or due to it acting at a late processing stage. However, long-term priors like the slow-motion prior act directly at a perceptual level (Sotiropoulos et al., 2011; Stocker & Simoncelli, 2004; Weiss et al., 2002). Therefore, the results here suggest that this pattern cannot be easily explained by high order, post-perceptual information acting to a greater extent on confidence. Another key difference between the high-level prior manipulations used previously and the low-level one used here is that participants were not aware of the influence of the slow-motion prior. This prior was naturally existing in participants, not manipulated in the lab. Further, the stimuli were designed such that participants would not even notice a decision bias forming from the prior, since it biased their decisions evenly towards blue and orange overall. This rules out the possibility that the strong effect of priors on confidence can be explained by participants realizing that they had this prior information and incorporating it post-decisionally, or by a

confidence bias induced by demand characteristics. The use of the confidence forced-choice paradigm further eliminates this possibility, reducing the chance of any report- or scale-based confidence biases (Mamassian, 2020; Mamassian & de Gardelle, 2022).

Here we find confidence to be biased towards information that is consistent with one's prior beliefs. This relates to other confirmatory confidence biases in the literature, such as the positive evidence bias in which confidence is biased towards information that confirms one's decision (Peters et al., 2017; Rollwage et al., 2020). Similarly to the positive evidence bias, we also find this to occur at a low level, and outside of the awareness of participants. Both of these confirmatory confidence biases may be adaptive for avoiding cognitive dissonance, and driving self-consistency (Caziot & Mamassian, 2021; Festinger, 1957). Also, although confirming prior-congruent information may seem maladaptive for maintaining good predictive models, having a positive epistemic feeling that reinforces prior-congruent information might actually help drive us to minimize prediction errors by encouraging us to seek information that will be in line with expectations (Friston et al., 2016; Wiese & Metzinger, 2017). The findings here may also relate to another confidence-performance dissociation that has been discussed in the literature, in which confidence is influenced by the visibility of a stimulus, over and above the total decision evidence (Rausch et al., 2018; Rausch & Zehetleitner, 2019). That work modeled confidence as a weighted combination of these two factors. In parallel with this, the prior-congruent evidence in the WPPCE model proposed here could reflect a visibility-like cue to confidence, if prior-congruent information increases perceptual strength. This is in line with work finding that prior-congruent stimuli enter conscious awareness earlier, and at lower thresholds (Aru et al., 2016; Eger et al., 2007; Esterman & Yantis, 2010; Melloni et al., 2011; Pinto et al., 2015; Stein & Peelen, 2015). Future work could test directly whether prior-congruent information from the slow-motion prior enhances the perceived motion, and whether it can promote earlier detection of stimuli.

Limitations

Although this work extends previous findings to what is often considered a different category of prior, it is still only one example of a long-term prior, and it is possible that others act in different ways. It would therefore be valuable to find ways to investigate other examples. In line with this, although we see the same confidence-performance dissociation with a long-term prior as was previously found with high-level priors, we cannot draw any conclusions about a shared mechanism of action. Future studies testing for correlations between the strengths of these dissociations across different priors, as well as neuroimaging work exploring this effect in

terms of neural implementation would be useful to clarify this. Finally, while it is clear that the slow-motion prior acts on perception, the precise way in which it impacts the phenomenal experience remains unclear. It seems possible that the prior-congruent information influences the perceptual strength of the motion, in turn impacting confidence, however we did not explicitly test this. The confidence forced-choice task has a benefit of reducing confounding confidence biases, but it also limits our insight into the subjective quality of the effect. Work testing whether information congruent with the slow-motion prior influences, for example, perceived speed, would be interesting. While it is important to find ways to investigate the impact of other long-term priors on confidence, it is also challenging to find ways to manipulate, or even quantify, these naturally formed low-level priors in participants. Here we provide a paradigm that tackles this by adjusting the decision axis to create conditions differing in the contribution of prior information to performance. This can provide a starting point with which to continue exploring the role of naturally formed perceptual priors at these different levels of processing.

Methods

The experiment was pre-registered (https://osf.io/3uc4k), and we respected the pre-registered plan unless stated otherwise.

Participants

We pre-registered that we would test 25 participants, ensuring at least 20 clean datasets that met the prespecified inclusion criteria, which most importantly included that they showed the basic effects of the prior – that the choice rates were biased towards the orthogonal motion direction in the Bias condition. We tested 25 participants and later excluded the data from one of them because they did not show the basic effect of the prior, leaving data from 24 participants (7 male, 17 female) included in our analyses. Participants were between 21 and 39 years of age (M=24.76, SD=4.28), reported to have normal or corrected-to-normal vision and were fluent in English. Participants were compensated with 10€ per hour and gave signed, informed consent before starting the experiment. The local ethics committee (Comité d'Ethique pour les Recherches en Santé) approved the study, which conformed to the Declaration of Helsinki.

Setup

The experiment was programmed using MATLAB (R2020b, Mathworks, Natick, MA, USA) and Psychtoolbox-3 (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). Participants sat in a

dark room with their head on a chin and forehead rest placed 57 cm away from a CRT monitor (Vision Mater Pro 454) with a display resolution of 1,920 x 1,200 (refresh rate=60Hz).

Procedure

ASA Procedure. At the beginning of the first session, after receiving verbal instruction about the stimuli, participants completed a short stimulus training consisting of 80 trials across five mini-blocks (16 trials each) in which the line motion stimulus was shown and they input their answer about whether the lines were moving towards the orange or blue region using the mouse. For the first mini-block, the true line motion direction was shown. In addition, the lines started at high contrast so that participants could clearly see and understand the stimuli, and got lower contrast in later blocks (Michelson contrasts: 10%, 10%, 3%, 3%, 3%), and their movement was displayed for a longer duration at first and got progressively shorter in each block (durations: 1000ms, 266ms, 133ms, 133ms, 133ms). After this training, participants completed a staircasing procedure in order for us to select the appropriate stimulus values in each condition for the full experiment, given their performance and their bias. The No-Bias condition consisted of only one setting, with the reference point exactly 90° away from both preferred, orthogonal line orientation directions. The Bias condition consisted of two settings, to capture the bias effect from the influence of the prior in both directions (a blue-bias and an orange-bias), by setting preferred, orthogonal line orientation either 35° towards the blue region or towards the orange region from reference (Fig. 1A). For each of these settings, we ran an accelerated stochastic approximation (ASA) staircasing procedure with 4 staircases, set to converge to two choice probability thresholds of 0.25 and 0.75. We then fit a cumulative normal distribution function ($\boldsymbol{\Phi}$) with a free noise (σ_{asa}) and bias (μ_{asa}) parameter to the data from each of these three separate ASA procedures:

$$P = \Phi\left(\theta, \mu_{asa}, \sigma_{asa}\right)$$

where P is the choice probability and θ is the stimulus value used, which was the angle between the reference and the motion direction (Fig. 1C). These fit psychometric functions were then used to select the stimulus values that corresponded to choice probabilities of 0.85 and 0.65 for the bias direction in each Bias condition setting, and choice probabilities of 0.85 and 0.65 for each different direction in the No-Bias condition. These stimulus values were used for the main task. This procedure took approximately 10 minutes.

Main Task. After completing the ASA procedure, participants received verbal instructions about the structure of the confidence forced-choice task, and completed another short training of 8 trials. Each trial of the main task consisted of two consecutive motion direction

decisions followed by a confidence choice. We used this criterion-free confidence forced-choice structure in order to reduce confidence biases that come from criterion placement or particular use of a confidence scale, which can make the interpretation of confidence results difficult (Mamassian, 2020; Mamassian & de Gardelle, 2022). The motion direction decisions worked in the same way as in the ASA procedure – participants viewed the matrix of moving lines at short duration (133ms) and low contrast (3.5% Michelson contrast) and then used the mouse to select whether the motion direction was towards the orange area or blue area by clicking on the corresponding colored quarter-ring around the display circle. In the confidence choice, participants used the left and right arrow keys to indicate whether they were more confident in the first or second interval, respectively (Fig. 1B). We used the chosen stimulus values (θ's, four per condition) from the ASA procedure to create 28 possible pairs, by using all possible pairings of these eight stimulus values except for ones with the identical stimulus value and condition twice. These 28 pairs formed the possible interval pairs and were counterbalanced across the main task in both possible interval orders. We fixed each block such that half the trials would have the line orientation rotated clockwise from vertical, and half counterclockwise from vertical, and amount of rotation was sampled randomly on each trial from the possible values: 10°, 20°, 30°, 40°, 50°, 60°, 70°, or 80°. After assigning this line orientation, the rotation of the reference and then line motion direction were set relative to the line orientation, and were dictated by the stimulus value pair – the condition dictated the placement of the reference and the θ dictated the line motion direction (Fig. 1A). We also fixed each block such that half the trials would have orange and half would have blue as the ring color counterclockwise to the reference. The order of trials was randomized within each block.

Participants completed 24 repetitions of each of the 28 stimulus pairs for a total of 672 trials, including 672 motion direction decisions in each condition (for a total of 1344 motion direction decisions) and 672 confidence choices. These took place across 12 blocks with 56 trials in each block. Four of these blocks were completed in the first session and the remaining eight took place in the second session. With the time taken for instructions and the ASA procedure in the first session, this resulted in each session taking approximately 1 hour.

Stimuli. The line motion stimulus was similar to the stimulus used in work by Sotiropolous et al., 2011. It consisted of a matrix of parallel line segments that moved rigidly, all in the same direction, and at a constant velocity. The line movement speed was 3° of visual angle per second and they were shown for a duration of 133ms. The matrix of lines was displayed through a circular mask in the center of the screen, which was 12° of visual angle in diameter. The line segments making up the matrix were 2.86° of visual angle, and were hence

much shorter than the circular mask, allowing many endpoints to be visible. This is critical because only the line ends provide the disambiguating information regarding the line velocity. The lines were 3.0 arcmin thick. The background was middle gray with a luminance of 0.5 cd/m^2 and the lines were lighter gray with a luminance of 0.536 cd/m^2, giving a Michelson contrast of 3.5%. For the orange/blue motion direction decision, the orange and blue choice regions were displayed by showing a colored half-ring around 180° of the circular mask, with 90° of that being blue and 90° of it being orange. This half-ring was visible for the duration of the line motion in order to avoid it appearing after and masking the stimulus. Participants then reported their decision by clicking the desired region of this half-ring. When participants moved the mouse close enough to one of the choice regions to select it, the orange or blue segment expanded slightly to highlight it, such that participants knew which they were selecting. This was helpful because the mouse cursor was invisible in order to avoid distracting participants during the stimulus display, or having a masking effect if it appeared immediately after. Before each stimulus, a green fixation point in the center of the screen was displayed for 500ms. The half-ring only appeared after this, with the matrix of lines, in order to avoid participants planning a strategy or fixating on the reference prior to the stimulus onset.

Analysis

As pre-registered, we removed any trials with reaction times that were longer than 8 seconds for any decision, or shorter than 100ms for any motion direction decision. We also removed trials that included a specific stimulus setting for which the basic bias effect was not present.

Our behavioral analysis testing the effect of condition on motion direction decisions was done using a logistic mixed-effects model with the ‘lme4’ package (Bates et al., 2015) in R (R Core Team, 2021). The test was two-tailed and used an alpha value of 0.05. We did not assess confidence choices using a regression approach because we prespecified that we would only do this if the first-order choice probabilities were adequately matched, which they were not. Additionally, we realized from further model exploration that the Bayesian optimal observer model does not predict equal confidence given matched first-order choice rates, so this analysis was not the best approach to assess deviations from the Bayesian optimal behavior in any case. Instead, and as planned given that first-order choice rates may not be perfectly matched, to examine confidence bias we used a non-parametric analysis following a similar approach to that in previous work (Toscani et al., 2021). We normalized the choice probability differences between the two conditions for each stimulus pair setting, excluding those with the same

condition in both intervals, by taking their z-score. For each of these normalized choice probability differences (cp_{diff}), we then computed the corresponding rate of choosing the Bias condition as the more confident one (*P("Bias")*) and fit a cumulative normal distribution function ($\mathbf{\Phi}$) with a free noise (σ_{cb}) and bias (μ_{cb}) parameter to this relationship:

$$P(\text{"}Bias\text{"}) = \Phi\left(cp_{diff}, \mu_{cb}, \sigma_{cb}\right)$$

Modeling

Quantifying Confidence Bias. In order to further quantify the confidence bias, before exploring the different process models that might explain it, we fit the *cfc-model* developed in previous work (Mamassian & de Gardelle, 2022). This model includes a confidence bias term, β, which generally captures the propensity to choose one condition as the more confident interval, after accounting for the stimulus strengths, first-order responses in each interval, and a potential interval bias (propensity to prefer a particular interval as the more confident one). In a single condition, confidence bias is captured as a parameter that scales the estimated sensory sensitivity. If this parameter is equal to 1, sensitivity is perfectly estimated. If this parameter is below 1, sensitivity is underestimated indicating underconfidence and if this parameter is above 1, sensitivity is overestimated indicating overconfidence. With two conditions, as in our case, the *cfc-model* fits β as the ratio of confidence biases between conditions with the No-Bias as the baseline, such that β=1 indicates no confidence bias, β>1 indicates a confidence bias favoring the Bias condition, and β<1 indicates a confidence bias favoring the No-Bias condition. In order to collapse the Bias condition into one psychometric function for the *cfc-model*, we projected the orange bias trials onto the blue bias space by taking:

$$\theta_{Orange \rightarrow Blue} = \mu_{No-Bias} - \theta_{Orange}$$

and

$$Response_{Orange \rightarrow Blue} = 1 - Response_{Orange}$$

We fit the model to the pooled data, using normalized stimulus intensity values (θ_{norm}) that were adjusted based on each participants' fit sensory noise ($\sigma_{No\text{-}Bias}$) and bias ($\mu_{No\text{-}Bias}$) from the No-Bias condition:

$$\theta_{norm} = \frac{\theta - \mu_{No-Bias}}{\sigma_{No-Bias}}$$

The model freely fit the sensory noise and criteria in each condition, along with the confidence parameters: confidence noise in each condition, confidence boost (reflecting additional information used in confidence) in each condition, interval bias, and the confidence bias

parameter, β. We did this for 100 bootstrapped runs of the model in order to extract the confidence intervals surrounding the estimate of β. We implemented and fit this model using the previously developed 'cfc' software (Mamassian, 2023). More on the details underlying the model and the other parameters fit can be found in the original work (Mamassian & de Gardelle, 2022).

Bayesian Decision Model. To capture the first-order motion direction decisions, we built a Bayesian observer model (Fig. 3A). It has a Gaussian mixture prior distribution of motion directions (in degrees away from the reference) with peaks at the two preferred, orthogonal motion directions (μ_{orth1} , μ_{orth2}) ,

$$prior \sim \left[0.5 * N\left(\mu_{orth1}, \sigma_P\right) + 0.5 * N\left(\mu_{orth2}, \sigma_P\right)\right]$$

where σ_P is the standard deviation of each Gaussian making up the mixture prior. In the No-Bias condition, μ_{orth1} and μ_{orth2} are at + and -90°. In the Bias condition, μ_{orth1} is 35° away from the reference in the bias direction, and μ_{orth2} is 145° away from the reference in the non-bias direction. The Gaussian likelihood is formed from the stimulus corrupted by sensory noise,

$$likelihood \sim N\left(r, \sigma_L\right)$$

where σ_L is the sensory noise and *r* is the internal signal caused by the stimulus. These combine to form the posterior distribution, and the perceived probability of the motion direction being in the orange (encoded as the negative direction) or blue (encoded as the positive direction) region is computed by the area under this posterior distribution across the entire orange or blue region, normalized:

$$P(Orange) = \frac{1}{N}\int_{-90}^{0}\left[0.5 * \varphi\left(x, \mu_{orth1}, \sigma_P\right) + 0.5 * \varphi\left(x, \mu_{orth2}, \sigma_P\right)\right] * \varphi\left(x, r, \sigma_L\right)dx \quad (1)$$

$$P(Blue) = \frac{1}{N}\int_{0}^{90}\left[0.5 * \varphi\left(x, \mu_{orth1}, \sigma_P\right) + 0.5 * \varphi\left(x, \mu_{orth2}, \sigma_P\right)\right] * \varphi\left(x, r, \sigma_L\right)dx \quad (2)$$

where *N* is equal to the total area under the posterior in the allowed decision region:

$$N = \int_{-90}^{90}\left[0.5 * \varphi\left(x, \mu_{orth1}, \sigma_P\right) + 0.5 * \varphi\left(x, \mu_{orth2}, \sigma_P\right)\right] * \varphi\left(x, r, \sigma_L\right)dx \quad (3)$$

The decision is then based on the max of these two values,

$$choice = max[\, P(Orange),\ P(Blue)\,]$$

We fit this model to the motion-direction decisions of each individual participant in order to find the σ_P and σ_L values that would best explain their first-order performance. However, in order to

deal with the computational demand of this we required a slight simplification. For the Bias condition, we considered the prior to be a single Gaussian centered around μ_{orth1}. Because the Gaussian centered around μ_{orth2} (+-145°) is so far outside of the allowed decision region (from -90° to +90°), this had a negligible impact on the model and made it possible to fit with the amount of data available to us. The ability of the model to still account well for first-order performance can be seen in Fig. 3C. We then used the fit σ_P and σ_L parameters for the rest of the confidence modeling described below.

Bayesian Optimal Confidence Model. We then considered the expected confidence patterns of an observer that follows the Bayesian Optimal confidence model. This model is in line with the Bayesian confidence model discussed in the literature in which confidence reflects the perceived posterior probability of being correct about a decision (Aitchison et al., 2015; Fleming & Daw, 2017; Kepecs & Mainen, 2012; Meyniel et al., 2015; Pouget et al., 2016; Sanders et al., 2016). In our case, confidence then corresponds to the perceived posterior probability that the orange/blue decision is correct, which is described above in equation (1) if orange is chosen, and (2) if blue is chosen. The confidence forced-choice after both intervals is then based on whichever of these two confidence values from the two intervals is higher:

$$cfc = max[\, conf1,\ conf2\,], \tag{4}$$

where the *conf* value is the perceived posterior probability correct. This model has no free parameters beyond the first-order sensitivity parameters, σ_P and σ_L.

Distance-To-Criterion Confidence Model. It is possible that, instead of computing the area under the posterior distribution as in the Bayesian Optimal confidence model, participants instead just base their confidence on one sample from the posterior. In this model, confidence reflects the distance between a sample from the posterior (r_c) and their criterion (*crit*):

$$conf = \left| r_c - crit \right| \tag{5}$$

High confidence decisions then occur when the internal posterior sample is far from the criterion, or is more clearly orange or blue. Again, the confidence forced-choice is then based on which of these confidence values is higher, as in equation (4). Like the Bayesian Optimal model, this model has no added free parameters for confidence.

Overweighted Prior Confidence Model. The Overweighted Prior confidence model reflects participants computing confidence as in the Bayesian Optimal confidence model above,

except overweighting the prior relative to the likelihood such that the prior has a stronger effect in confidence compared to the first-order decisions. The relative overweighting of the prior is implemented by an underestimation of the sensory precision of the likelihood through a scalar weighting parameter *w*, such that the estimated posterior probabilities of each choice at the confidence level are:

$$P(Orange) = \frac{1}{N}\int_{-90}^{0}\left[0.5 * \varphi\left(x, \mu_{orth1}, \sigma_P\right) + 0.5 * \varphi\left(x, \mu_{orth2}, \sigma_P\right)\right] * \varphi\left(x, r, w * \sigma_L\right)dx \quad (6)$$

$$P(Blue) = \frac{1}{N}\int_{0}^{90}\left[0.5 * \varphi\left(x, \mu_{orth1}, \sigma_P\right) + 0.5 * \varphi\left(x, \mu_{orth2}, \sigma_P\right)\right] * \varphi\left(x, r, w * \sigma_L\right)dx \quad (7)$$

and *N* is:

$$N = \int_{-90}^{90}\left[0.5 * \varphi\left(x, \mu_{orth1}, \sigma_P\right) + 0.5 * \varphi\left(x, \mu_{orth2}, \sigma_P\right)\right] * \varphi\left(x, r, w * \sigma_L\right)dx \quad (8)$$

If *w*=1, the sensory precision is correctly estimated in the confidence computation and the prior and likelihood are weighted in the same way as in the first-order decisions. This, additionally, would be identical to the Bayesian Optimal confidence model. If *w*<1, the sensory noise is underestimated, and the prior is underweighted relative to the likelihood. On the contrary, if *w*>1, the sensory noise is overestimated, and the prior is overweighted relative to the likelihood. In this model, the weighting parameter *w* was fit as a free parameter.

Prior-Congruent Evidence (PCE) Confidence Model. In the PCE confidence model, confidence is not based on the Bayesian posterior distribution but is instead based on the amount of prior-congruent information in the stimulus. For these line motion stimuli, the prior-congruent information is the information in the preferred, orthogonal motion direction (Fig. 6A). So, the confidence is proportional to the length of the vector component in this orthogonal direction,

$$conf = PCE = \left|radius * cos(\theta_r - \theta_P)\right| \quad (9)$$

where θ_r is the angle between the reference and the motion direction of the internal signal generated by the stimulus; θ_P is the angle between the reference and the preferred, orthogonal direction; and *radius* is the radius of the stimulus circle, which we set to 1 for simplicity. The confidence choice is then based on which of these *PCE* values is larger, or which of the two intervals had more prior-congruent evidence, independently of the posterior evidence and the first-order responses.

Weighted Posterior and Prior-Congruent Evidence (WPPCE) Confidence Model. The WPPCE confidence model suggests that the prior-congruent evidence described above serves as an additional signal, but not the sole signal, to confidence. In this model, confidence is based on the weighted combination of this *PCE* and the Bayesian posterior evidence (PostE), weighted according to a weighting parameter *α*,

$$conf = (1 - \alpha) * PostE + \alpha * PCE \quad (10)$$

where the *PCE* term is computed as in equation (9), and the *PostE* term is computed as in equation (1) or (2), depending on the response. Larger values of *α* indicate more use of the prior-congruent evidence in confidence, and smaller values of *α* indicate more use of the posterior evidence. When *α*=0, this model is identical to the Bayesian Optimal confidence model and when *α*=1, this model is identical to the PCE confidence model. The weighting term, *α*, was a free parameter.

Model-Fitting and Comparison. The confidence models with additional free parameters for confidence – the Overweighted Prior model and the WPPCE model – were fit using a maximum likelihood estimation (MLE) approach to find the *w* and *α* parameter values respectively that could best explain the confidence forced-choice data. We used a simulation-based approach to approximate the likelihood, which was analytically intractable. This was also done to compute the likelihood of the models without additional free parameters, to assess model fit. Model fitting was done in R using the 'stats4' package. In fitting the Overweighted Prior model, we did an initial grid search across values of *w* from 0.1 to 4, in steps of 0.1, and used the *w* value that corresponded with the highest likelihood as the start value of the MLE procedure. This was done to avoid convergence issues. For a model comparison, we compared AIC between the models, computed according to

$$AIC = 2 * k - 2 * log(\widehat{L})$$

where *k* is the number of free parameters and $\widehat{L}$ is the maximized value of the likelihood function. We considered a model to be conclusively better if its AIC was lower by at least 2, following convention in the literature. We ensured the models to be distinguishable from one another in a model recovery analysis (Supplementary Information; Fig. S1).

Acknowledgments

MC's work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—337619223 / RTG2386. This work was supported by a Freigeist Fellowship from

the Volkswagen Foundation (grant number 91620) to EF. EF and PM were supported by a European Commission Doctoral Network grant “CODE” (EC MSCA-101119647). The funders had no role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript.

References

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the “light-from-above” prior. *Nature Neuroscience*, *7*(10), Article 10. https://doi.org/10.1038/nn1312

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making. *PLOS Computational Biology*, *11*(10), e1004519. https://doi.org/10.1371/journal.pcbi.1004519

Aru, J., Rutiku, R., Wibral, M., Singer, W., & Melloni, L. (2016). Early effects of previous experience on conscious perception. *Neuroscience of Consciousness*, *2016*(1), niw004. https://doi.org/10.1093/nc/niw004

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), Article 1. https://doi.org/10.18637/jss.v067.i01

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Caziot, B., & Mamassian, P. (2021). Perceptual confidence judgments reflect self-consistency. *Journal of Vision*, *21*(12), 8. https://doi.org/10.1167/jov.21.12.8

Constant, M., Pereira, M., Faivre, N., & Filevich, E. (2023). Prior information differentially affects discrimination decisions and subjective confidence reports. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-41112-0

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How Do Expectations Shape Perception? *Trends in Cognitive Sciences*, *22*(9), 764–779. https://doi.org/10.1016/j.tics.2018.06.002

Eger, E., Henson, R. N., Driver, J., & Dolan, R. J. (2007). Mechanisms of top-down facilitation in perception of visual objects studied by FMRI. *Cerebral Cortex (New York, N.Y.: 1991)*, *17*(9), 2123–2133. https://doi.org/10.1093/cercor/bhl119

Esterman, M., & Yantis, S. (2010). Perceptual expectation evokes category-selective cortical activity. *Cerebral Cortex (New York, N.Y.: 1991)*, *20*(5), 1245–1253. https://doi.org/10.1093/cercor/bhp188

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press.

Fleming, S. M., & Daw, N. D. (2017). Self-Evaluation of Decision-Making: A General Bayesian Framework for Metacognitive Computation. *Psychological Review*, *124*(1), 91–114. https://doi.org/10.1037/rev0000045

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O’Doherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, *68*, 862–879. https://doi.org/10.1016/j.neubiorev.2016.06.022

Gerardin, P., Kourtzi, Z., & Mamassian, P. (2010). Prior knowledge of illumination for 3D

perception in the human brain. *Proceedings of the National Academy of Sciences*, *107*(37), 16309–16314. https://doi.org/10.1073/pnas.1006285107

Hardstone, R., Zhu, M., Flinker, A., Melloni, L., Devore, S., Friedman, D., Dugan, P., Doyle, W. K., Devinsky, O., & He, B. J. (2021). Long-term priors influence visual perception through recruitment of long-range feedback. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-26544-w

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1322–1337. https://doi.org/10.1098/rstb.2012.0037

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object Perception as Bayesian Inference. *Annual Review of Psychology*, *55*(1), 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005

Kleiner, M., Brainard, D. H., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, *36*, 1–16. https://doi.org/10.1068/v070821

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian Inference*. Cambridge University Press. https://doi.org/10.1017/CBO9780511984037

Locke, S. M., Gaffin-Cahn, E., Hosseinizaveh, N., Mamassian, P., & Landy, M. S. (2020). Priors and payoffs in confidence judgments. *Attention, Perception, & Psychophysics*, *82*(6), 3158–3175. https://doi.org/10.3758/s13414-020-02018-x

Locke, S. M., Landy, M. S., & Mamassian, P. (2022). Suprathreshold perceptual decisions constrain models of confidence. *PLOS Computational Biology*, *18*(7), e1010318. https://doi.org/10.1371/journal.pcbi.1010318

Mamassian, P. (2020). Confidence Forced-Choice and Other Metaperceptual Tasks. *Perception*, *49*(6), 616–635. https://doi.org/10.1177/0301006620928010

Mamassian, P. (2023). *Cfc* (1.0.0) [Computer software]. https://doi.org/10.5281/zenodo.7791435

Mamassian, P., & de Gardelle, V. (2022). Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychological Review*, *129*(5), 976–998. https://doi.org/10.1037/rev0000312

Mamassian, P., & Goutcher, R. (2001). Prior knowledge on the illumination position. *Cognition*, *81*(1), B1-9. https://doi.org/10.1016/s0010-0277(01)00116-0

Mamassian, P., Landy, M., & Maloney, L. T. (2002). Bayesian Modelling of Visual Perception. In R. P. N. Rao, B. A. Olshausen, & M. S. Lewicki (Eds.), *Probabilistic Models of the Brain* (pp. 13–36). The MIT Press. https://doi.org/10.7551/mitpress/5583.003.0005

McNamara, J. M., Green, R. F., & Olsson, O. (2006). Bayes' theorem and its applications in

animal behaviour. *Oikos*, *112*(2), 243–251. https://doi.org/10.1111/j.0030-1299.2006.14228.x

Melloni, L., Schwiedrzik, C. M., Müller, N., Rodriguez, E., & Singer, W. (2011). Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(4), 1386–1396. https://doi.org/10.1523/JNEUROSCI.4570-10.2011

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron*, *88*(1), 78–92. https://doi.org/10.1016/j.neuron.2015.09.039

Montagnini, A., Mamassian, P., Perrinet, L., Castet, E., & Masson, G. S. (2007). Bayesian modeling of dynamic motion integration. *Journal of Physiology, Paris*, *101*(1–3), 64–77. https://doi.org/10.1016/j.jphysparis.2007.10.013

Olawole-Scott, H., & Yon, D. (2023). Expectations about precision bias metacognition and awareness. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0001371

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

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

Pinto, Y., van Gaal, S., de Lange, F. P., Lamme, V. A. F., & Seth, A. K. (2015). Expectations accelerate entry of visual stimuli into awareness. *Journal of Vision*, *15*(8), 13. https://doi.org/10.1167/15.8.13

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

R Core Team. (2021). *R: A language and environment for statistical computing.* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

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. https://doi.org/10.3758/s13414-017-1431-5

Rausch, M., & Zehetleitner, M. (2019). The folded X-pattern is not necessarily a statistical signature of decision confidence. *PLOS Computational Biology*, *15*, e1007456.

https://doi.org/10.1371/journal.pcbi.1007456

Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-16278-6

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

Schmack, K., Bosc, M., Ott, T., Sturgill, J. F., & Kepecs, A. (2021). Striatal dopamine mediates hallucination-like perception in mice. *Science*, *372*(6537), eabf4740. https://doi.org/10.1126/science.abf4740

Seriès, P., & Seitz, A. R. (2013). Learning what to expect (in visual perception). *Frontiers in Human Neuroscience*, *7*, 668. https://doi.org/10.3389/fnhum.2013.00668

Sherman, M., & Seth, A. (2021). *Effects of expected task difficulty on metacognitive confidence and multitasking*. PsyArXiv. https://doi.org/10.31234/osf.io/3gfp2

Sherman, M. T., Kanai, R., Seth, A. K., & VanRullen, R. (2016). Rhythmic Influence of Top–Down Perceptual Priors in the Phase of Prestimulus Occipital Alpha Oscillations. *Journal of Cognitive Neuroscience*, *28*(9), 1318–1330. https://doi.org/10.1162/jocn_a_00973

Sherman, M. T., Seth, A. K., Barrett, A. B., & Kanai, R. (2015). Prior expectations facilitate metacognition for perceptual decision. *Consciousness and Cognition*, *35*, 53–65. https://doi.org/10.1016/j.concog.2015.04.015

Sotiropoulos, G., Seitz, A. R., & Seriès, P. (2011). Changing expectations about speed alters perceived motion direction. *Current Biology: CB*, *21*(21), R883-884. https://doi.org/10.1016/j.cub.2011.09.013

Stein, T., & Peelen, M. V. (2015). Content-specific expectations enhance stimulus detectability by increasing perceptual sensitivity. *Journal of Experimental Psychology: General*, *144*(6), 1089–1104. https://doi.org/10.1037/xge0000109

Stocker, A. A., & Simoncelli, E. (2004). Constraining a Bayesian Model of Human Visual Speed Perception. *Advances in Neural Information Processing Systems*, *17*. https://proceedings.neurips.cc/paper/2004/hash/852c44ddce7e0c7e4c64d86147300831-Abstract.html

Stocker, A. A., & Simoncelli, E. P. (2007). A Bayesian Model of Conditioned Perception. *Advances in Neural Information Processing Systems*, *2007*, 1409–1416.

Teufel, C., & Fletcher, P. C. (2020). Forms of prediction in the nervous system. *Nature Reviews*

Neuroscience, *21*(4), Article 4. https://doi.org/10.1038/s41583-020-0275-5

The MathWorks Inc. (2020). MATLAB version: 9.9.0 (R2020b), Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com

Toscani, M., Mamassian, P., & Valsecchi, M. (2021). Underconfidence in peripheral vision. *Journal of Vision*, *21*(6), 2. https://doi.org/10.1167/jov.21.6.2

Van Marcke, H., Denmat, P. L., Verguts, T., & Desender, K. (2024). Manipulating Prior Beliefs Causally Induces Under- and Overconfidence. *Psychological Science*, *35*(4), 358–375. https://doi.org/10.1177/09567976241231572

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), Article 6. https://doi.org/10.1038/nn0602-858

Wiese, W., & Metzinger, T. (2017). Vanilla PP for Philosophers: A Primer on Predictive Processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308. https://doi.org/10.1016/j.tics.2006.05.002

Zellner, A. (1995). Bayesian and non-Bayesian approaches to statistical inference and decision-making. *Journal of Computational and Applied Mathematics*, *64*(1), 3–10. https://doi.org/10.1016/0377-0427(95)00002-X

Zhang, L.-Q., & Stocker, A. A. (2022). Prior Expectations in Visual Speed Perception Predict Encoding Characteristics of Neurons in Area MT. *Journal of Neuroscience*, *42*(14), 2951–2962. https://doi.org/10.1523/JNEUROSCI.1920-21.2022

Supplementary Information

Model Recovery Analysis

We ran a model recovery analysis to ensure that the models were adequately distinguishable in which we simulated data from each model and then checked that the correct model could be recovered. For each of 10 repetitions, we simulated 672 trials of data (equivalent to a single participant) from each model, using representative stimulus values by taking the mean stimulus intensity across participants for each expected performance setting. We also used representative sensitivity values by taking the median fit first-order noise parameters (σ_L and σ_P) across participants, and for the models with additional free parameters for confidence, we used the best fitting values from the group fits: w=1.54 for the Overweighted Prior Model and α=0.34 for the WPPCE Model. Models were compared using AIC, with the winning model captured as that with the lowest AIC value. Results of this analysis are shown in Fig. S1, revealing for each true generative model the proportion of repetitions for which each competing model was recovered as the winning one. This shows that there were only two repetitions in which data were generated from the Bayesian Optimal Observer model and either the WPPCE or OWP was recovered as the best model, rather than the correct generative model. This is also not completely unexpected, as both the WPPCE and OWP models have certain parameter settings (α=0 and w=1, respectively) for which they are indistinguishable from the Bayesian Optimal Observer model. Critically, when data is generated from the WPPCE model with the additional influence of the prior-congruent evidence (α>0), this model is always recovered correctly, suggesting the conclusions from our model comparisons to be reliable.

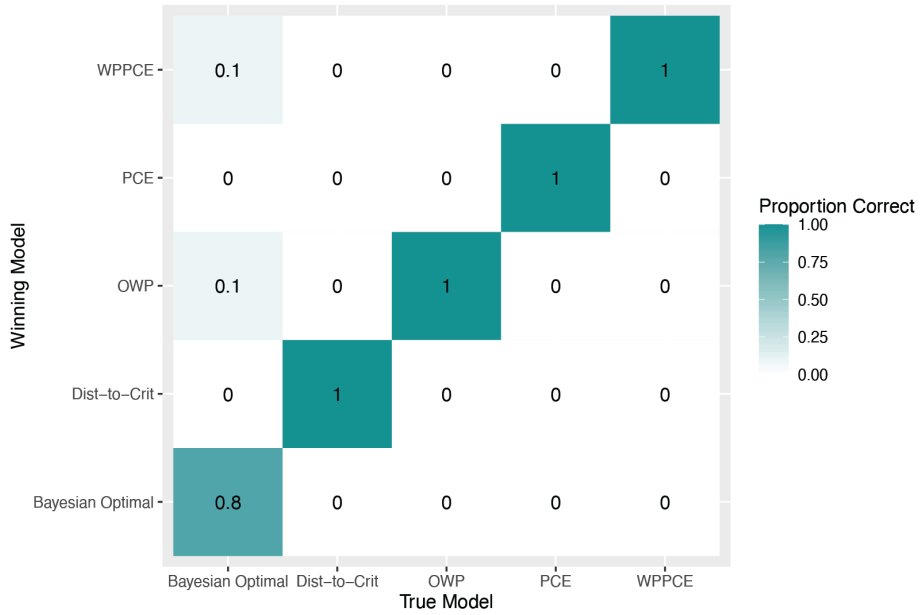

Figure S1. Model Recovery Analysis. We tested model recovery by simulating data from each model, and then comparing each competing model's fit to that data using AIC. Columns reflect the true generative model and rows reflect the recovered model. Cell color captures the proportion out of the 10 repetitions for which each competing model (row) was recovered as the winning one, given the true generative model (column). Cells diverging from the diagonal therefore indicate cases for which the incorrect model was recovered as the best fitting one.