bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introspective inference counteracts perceptual distortion

Andra Mihali,[1,2] Marianne Broeker,[1,2,3,4], Florian Ragalmuto [1,2,5], Guillermo Horga[1,2]

[1]New York State Psychiatric Institute, New York, NY, USA

[2]Columbia University, Department of Psychiatry, New York, NY, USA

[3]Columbia University, Teachers College, New York, NY, USA

[4]University of Oxford, Department of Experimental Psychology, Oxford, UK

[5] Vrije Universiteit, Department of Cognitive Neuroscience, Amsterdam, NL

September 29, 2022

Abstract

Introspective agents can recognize the extent to which their internal perceptual experiences deviate from the actual states of the external world. This ability, also known as insight, is critically required for reality testing and is impaired in psychosis, yet very little is known about its cognitive underpinnings. We developed a Bayesian modeling framework and a novel psychophysics paradigm to quantitatively characterize this type of insight while participants experienced a motion after-effect illusion. Participants could compensate for the illusion when judging the actual direction of a motion stimulus. Furthermore, confidence, reaction-time, and pupil-dilation data all showed signatures consistent with inferential adjustments in the Bayesian insight model. Our results suggest that people can question the veracity of what they see and make insightful inferences that incorporate introspective knowledge about internal distortions.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Successfully navigating the modern world requires questioning the validity of sensory information - testing whether it conforms to reality rather than simply taking it at face value. In other words, one cannot always believe what one sees. This is apparent in considering "deepfakes", where objective sensory information is deliberately manufactured in such lifelike ways that viewers often believe it is real. The corollary is that to avoid harboring false beliefs, and acting erroneously on their basis, the presence of potentially deceiving sensory information needs to be met with a healthy degree of skepticism in one's own senses. The ability to perform such "reality testing" is characteristically impaired in psychosis [1,2]. Although perceptual distortions, including hallucinations, are common in a variety of neuropsychiatric conditions as well as in some non-clinical groups [3], psychotic disorders such as schizophrenia are classically distinguished by an inability to question the reality of distorted percepts, a concept referred to as impaired insight [4]. Impaired insight can lead to false beliefs and drive seemingly erratic behaviors, including poor treatment adherence [5, 6]. Alterations in reality testing also have important legal ramifications given their impact on determinations of responsibility for criminal actions. Despite the broad societal implications of reality testing, the cognitive mechanisms underlying this ability are poorly understood [4].

Progress in this area has been hindered by the lack of a formal explanatory framework and experimental paradigms suitable to study reality testing. Previous work related to reality testing has broadly used two main approaches: source (self versus other) memory tasks in the context of semantic association and sentence completion [7, 8] and, more recently, imagery tasks in the context of perceptual decision-making [9–12]. The first approach relies on episodic memory and does not capture in-the-moment reality testing; the second approach captures in-the-moment processes and leverages signal-detection theory but relies on individuals' imagery ability, thus limiting experimental control and translatability. Furthermore, while imagery is relevant to some forms of reality testing, reality testing is a broader construct [11] and its impairments need not involve altered imagery. Here, we focus on a form of reality testing that we refer to as 'perceptual insight': the in-the-moment process of incorporating introspective knowledge about distortions of internal percepts to effectively infer the actual state of the world. We propose a formal model of perceptual insight building on Bayesian theory and present a novel perceptual-insight paradigm that minimizes memory confounds and enhances experimental control over previous designs.

We reasoned that probing perceptual insight would first require experimentally inducing a sufficiently strong perceptual distortion - a discrepancy between an objective stimulus and its corresponding percept - and then probing beliefs about the true state of varying objective stimuli. To do this, we used the motion after-effect (MAE), a well-characterized class of illusions that includes strong illusory percepts of (seemingly veridical) complex motion induced by prolonged viewing of an adaptor motion stimulus. We deemed the complex MAE particularly suitable for studying perceptual insight because explicit knowledge about the illusion seemed to allow accurate inferences on the actual state of objective stimuli while their subjective experience was distorted. Augmenting classic MAE methods with novel psychophysics tools, computational modeling, and pupillometry, we thus set out to validate an explanatory framework for human perceptual insight.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Formal model of perceptual insight

The construct of perceptual insight we focus on, alongside related notions of introspection, implies an ability to make judgments incorporating knowledge that the internal representation of the external states of the world may not accurately reflect these states – i.e., that internal experiences may not match reality. In particular, it implies incorporating knowledge that internal representations may be systematically influenced by factors other than external states (e.g., knowing that a voice one hears is "in one's head" and does not correspond to a speaker in the outside world).

To formulate this construct, we first consider a standard model of perceptual decision-making [13,14]. Broadly, this model captures how an agent observing a stimulus s makes a decision about the stimulus, for instance the category C it belongs to. This process involves an early sensory-encoding stage, in which the external stimulus s is encoded into a neural measurement or internal sensory representation x (Fig. 1A, left). Based on x, which is corrupted by noise, at an intermediate stage the agent infers aspects of the external stimulus such as its category C. Finally, at a late stage, the agent makes a choice $\hat{C}$ with a certain level of confidence q. A Bayesian observer formally solves this problem by first inferring the posterior probability of C at the intermediate stage, combining its likelihood and prior probability as $p(C|x) \propto p(x|C)\ p(C)$, and computing a perceptual-decision variable d consisting of the log-posterior ratio of the two possible options. The observer then chooses the category with the highest posterior probability at the late stage by comparing d to a threshold k_{choice}, with a value of 0 in the unbiased case. The observer further generates a confidence response q by comparing the posterior probability of the chosen response $p(C = \hat{C}|x)$ to a threshold $k_{\text{confidence}}$ [15–17] (see Methods S2.1).

Now we consider a scenario where the internal representation x is additionally influenced by a distortion factor A and is no longer solely a function of the external stimulus s. Factor A (e.g., a psychotomimetic substance) can distort the internal representation x, for instance offseting x from s (Fig. 1A, center). Similar to the standard case, an insightful Bayesian observer can infer the posterior probability $p(C|x)$ but will require incorporating knowledge about the distortion factor A, specifically about the influence of A on x and the possible values of A, as $p(C|x) \propto \int p(x|C, A)\ p(C)\ p(A)\ dA$ (see Methods S2). This agent has perceptual insight because it knows its internal representation x can be influenced by factors other than external states (Fig. 1B, right). By incorporating this knowledge at an intermediate inferential stage thus shifting the perceptual-decision variable d, the insightful agent can compensate for factor A and make accurate decisions about category C using an unbiased response rule - i.e., choosing the option with the highest posterior probability by comparing d to a k_{choice} of 0 (Fig. 1C, right).

We must distinguish this insightful agent (Fig. 1C, right) from one that behaves similarly under factor A but lacks perceptual insight (Fig. 1C, center). Without explicitly incorporating knowledge about factor A at the intermediate inferential stage, this alternative agent may still compensate for factor A at a late stage by adjusting its response rule, i.e., via a response bias (as if it knew its decisions were inaccurate but not the underlying cause). This agent would infer $p(C|x)$ as $p(x|C)\ p(C)$ - that is, not incorporating knowledge about factor A into the perceptual-decision variable d - and would only change its responses by using a k_{choice} different than 0 (Fig. 1C, center). This is reminiscent of patients with impaired insight who nonetheless learn to report skepticism about their distorted percepts (such as hallucinations) to derive some benefit.

Here, we set out to test whether human observers experiencing internal distortions in perception demonstrate perceptual insight or instead exhibit a late compensation in the form of a response bias. Because the critical difference between these possibilities is a distinct shift in the perceptual-decision variable d under perceptual insight (Fig. 1D, center versus right), through which the insightful agent

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

incorporates introspective knowledge about the internal distortion into its posterior beliefs, we placed particular emphasis on behavioral and physiology measures shown to reflect this variable d. More generally, our study is distinct from previous extensive work showing that observers integrate external stimuli [18–21] and external feedback [14, 22, 23] into perceptual decisions in line with Bayesian theory, in that it is, to our knowledge, the first to examine integration of evidence from external stimuli with introspective knowledge about an internal distortion in the absence of external feedback. Furthermore, previous work has shown that perceptual-decision changes at different stages can manifest as distinct patterns of choice, reaction times (RT), confidence reports [24, 25], and pupil dilation [26], which together can help differentiate shifts in the perceptual-decision variable d such as those we specifically hypothesize under perceptual insight.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

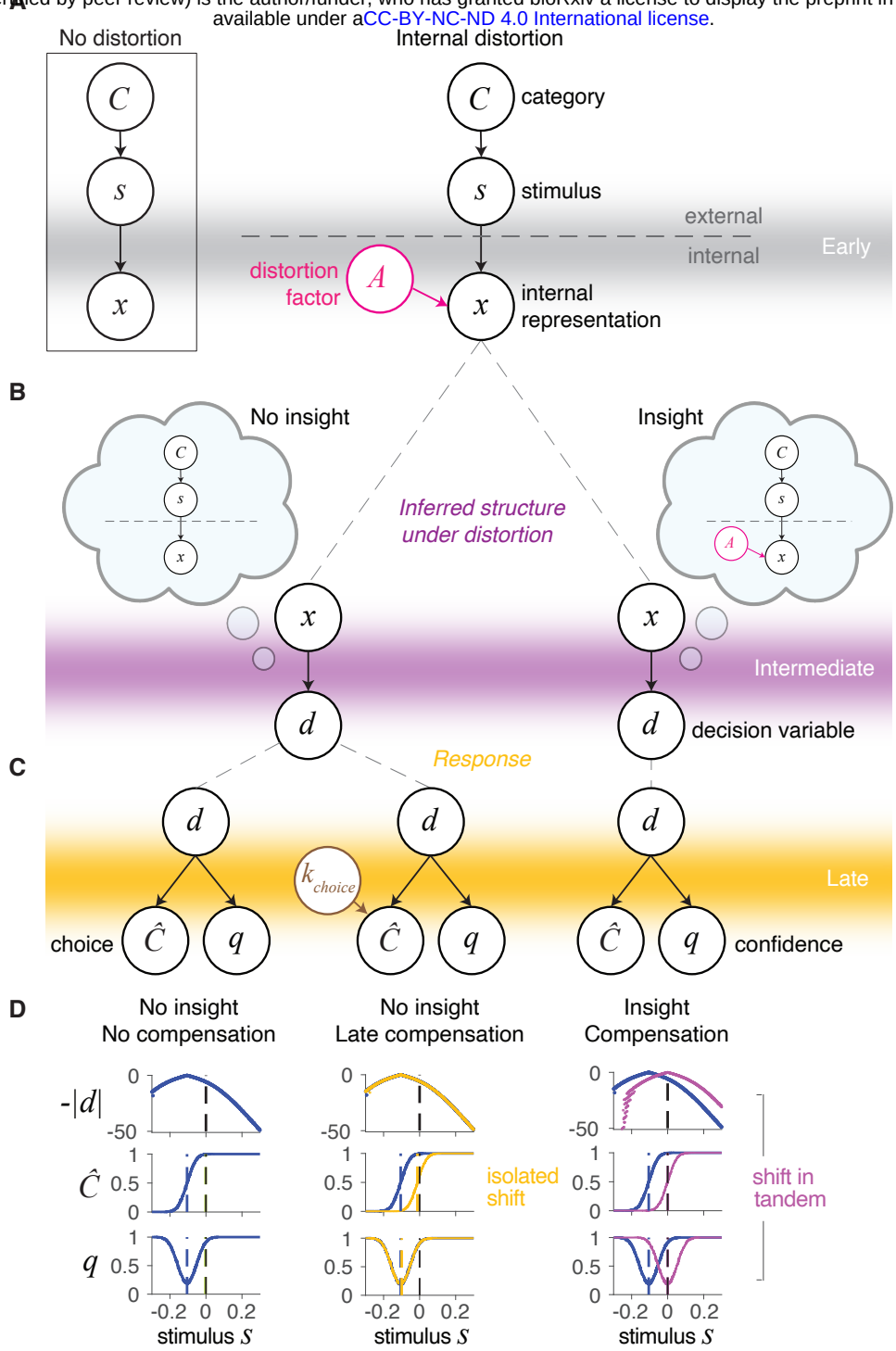

Figure 1: **Perceptual-decision framework for perceptual insight. A)** Generative model of the early encoding stage with internal distortion due to factor A (center) and with no distortion (left box). **B)** At the intermediate stage, an observer's inference under distortion can either take into account the actual generative model and thus demonstrate insight (right cloud) or wrongly assume the generative model without distortion and thus lack insight (left cloud). **C)** Response - choice ($\hat{C}$) and confidence (q) - generation from the decision variable d at the late stage. **D)** Choice and confidence patterns with insight (right) and without insight (center and left) are distinct. Compensation for the distortion through perceptual insight (right) leads to shifts in tandem of the choice and confidence curves. Late compensation with no insight (center) leads to isolated shifts in the choice curves.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Experimental paradigm of perceptual insight

To produce a strong perceptual distortion, we induced a MAE illusion of complex motion using a counterclockwise-rotating Archimedean spiral, based on the ability of complex motion (e.g., rotational) to induce stronger adaptation than simple motion (e.g., translational) [27, 28]. We measured MAE strength via the nulling method [29] as the shift in the bias of the psychometric curve, i.e., the bias induced by a first moving spiral (the adaptor) on the observer's judgment of a second spiral moving at different speeds and directions. The overall paradigm (Fig. 2) had a two-by-two design, with two types of conditions, 'Adapt' (rotating first spiral acting as adaptor) and 'No-Adapt' (static first spiral acting as control), and two types of responses required upon observing the second spiral (test stimulus), 'See' and 'Believe'. 'See' responses required participants to report the perceived direction of motion of the test stimulus (i.e., whether they saw the second spiral moving counterclockwise [left] or clockwise [right]); 'Believe' responses instead required them to report the inferred true direction of motion of the test stimulus (i.e., whether participants thought that the second spiral was actually moving counterclockwise or clockwise). Adapt blocks started with an extended exposure to an adaptor spiral (as in [30]) followed by trial-by-trial 3-second exposures to the adaptor spiral before each stimulus, and adaptor spirals always rotated counterclockwise at a constant speed. After each See or Believe response, participants also reported how sure they were about their choice (high or low confidence). Participants received no feedback on their responses during the task. Critically, participants had explicit knowledge of the MAE and its illusory nature, acquired through detailed instructions and practice and individually demonstrated in MAE strength estimates (Methods S1.3.1, Fig. S1). We measured MAE compensation as a relative correction of the Adapt-See bias (i.e., the MAE illusion) in the Adapt-Believe condition and thus a candidate measure for perceptual insight. This corrective shift may in principle reflect participants' knowledge of the direction and strength of the illusion since both conditions had matched stimuli and only differed in their required responses.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

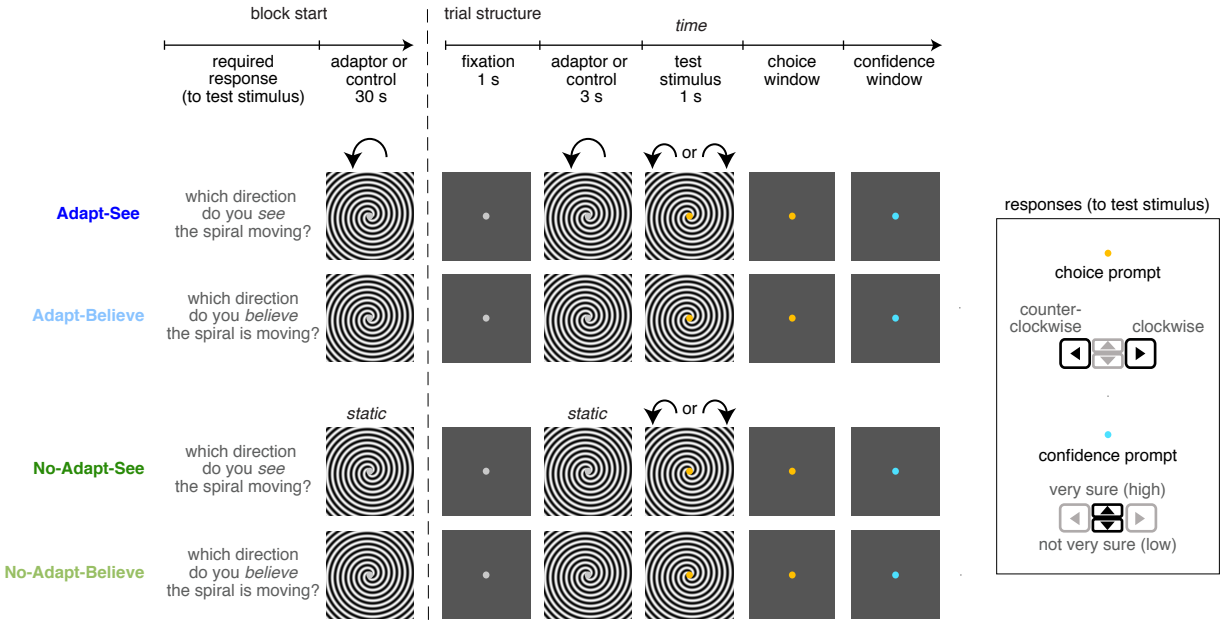

Figure 2: **Perceptual-insight task.** The four task conditions differing on the presence of a rotating spiral adaptor or static control spiral (**Adapt**, top blue, versus **No-Adapt**, bottom green) and required responses (**See**, dark colors, versus **Believe**, light colors) are depicted. Details (e.g., prompt colors and response-window duration) correspond to Experiment 2 (see Methods S1.3 for minor differences in Experiment 1). Blocks start with a reminder of required responses followed by a 30-s spiral rotating counterclockwise at constant speed or a static control spiral. Trials start with a fixation screen followed by the first spiral (rotating adaptor at constant speed counterclockwise or static control) and then a second spiral (test stimulus) of variable speed. A binary left/right choice is then prompted about motion direction (for clockwise or counterclockwise motion in the test stimulus, respectively, as seen or believed depending on the required response), followed by an up/down confidence response.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Participants compensate for distorted perception

In a first experiment, participants experiencing the MAE illusion showed evidence of compensation (Fig. 3, top). Observers' responses tracked stimulus strength derived from a staircase procedure (Methods S1.2; Fig. 3A, top) and were well described by psychometric curves (Gaussian cumulative density functions) with condition-specific bias μ and noise σ, and a shared lapse parameter λ (Fig. 3B, see Methods S1.7). The MAE appeared as a leftward shift in the psychometric curves in Adapt-See relative to the control No-Adapt-See condition, indicating a bias towards perceiving clockwise motion (difference in μ bias parameter: $z = 3.247, p = 0.0012$, Wilcoxon signed-rank test). Critically, relative to Adapt-See, the psychometric curves for Adapt-Believe showed a rightward corrective shift, indicating a compensation for the MAE ($z = 2.52, p = 0.012$; Fig. 3C, bottom; note that throughout this MAE compensation index is normalized by $|\mu_{\text{Adapt-See}}|$).

Despite evidence for MAE compensation in Experiment 1, we noted substantial interindividual variability and could not rule out the possibility that participants solved the Adapt-Believe condition by actively eluding the adaptor to minimize its influence (e.g., subtly looking away despite the instructions).To control for this, Experiment 2 used eye tracking to enforce fixation and prevent blinks for a 8-s time window encompassing the fixation, adaptor and stimulus periods and 3 s post-stimulus. We also sampled stimuli from a uniform distribution (informed by Experiment 1) to improve interpretability and facilitate modeling [31]. All results from Experiment 1 were replicated and more evident in Experiment 2 (Fig. 3, bottom), which generally produced higher-quality data: participants experienced the MAE and consistently compensated for it in Adapt-Believe (Fig. 3C, bottom; $z = 4.11, p = 3.99 \cdot 10^{-5}$, Wilcoxon signed-rank test). Furthermore, unlike in Experiment 1, the noise-parameters σ were comparable between Adapt-See and Adapt-Believe in Experiment 2 ($z = -1.41, p = 0.16$), and so were the lapse parameters λ in a model with condition-specific λ (not shown, $z = -1.72, p = 0.08$), suggesting the two critical task conditions were directly comparable and only differed consistently in their bias μ. Additionally, in Experiment 2, participants tended to exhibit a systematic overcompensation as psychometric curves in Adapt-Believe were often shifted rightward beyond the control condition (Fig. 3A, bottom and Fig. 3C, bottom, $z = 3.94, p = 7.99 \cdot 10^{-5}$, Wilcoxon signed-rank test for MAE compensation above 1). In sum, our participants experienced illusory motion and were able to either report the distorted percept (in Adapt-See) or discount it and compensate for it (in Adapt-Believe). Critically, they responded differently in Adapt-See and Adapt-Believe despite experiencing identical (adaptor and test) sensory stimuli across these two conditions.

MAE compensation could in principle reflect a more general compensatory strategy not due to insight gained from training and applied during the task. However, we empirically confirmed via MAE strength estimation that participants knew about the illusion and expected to have MAE roughly consistent in magnitude with the observed MAE effect during Adapt-See (Supplementary Fig. S1). Furthermore, because participants received no feedback during the task, it is unlikely that they could compensate for the MAE via trial-and-error learning.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

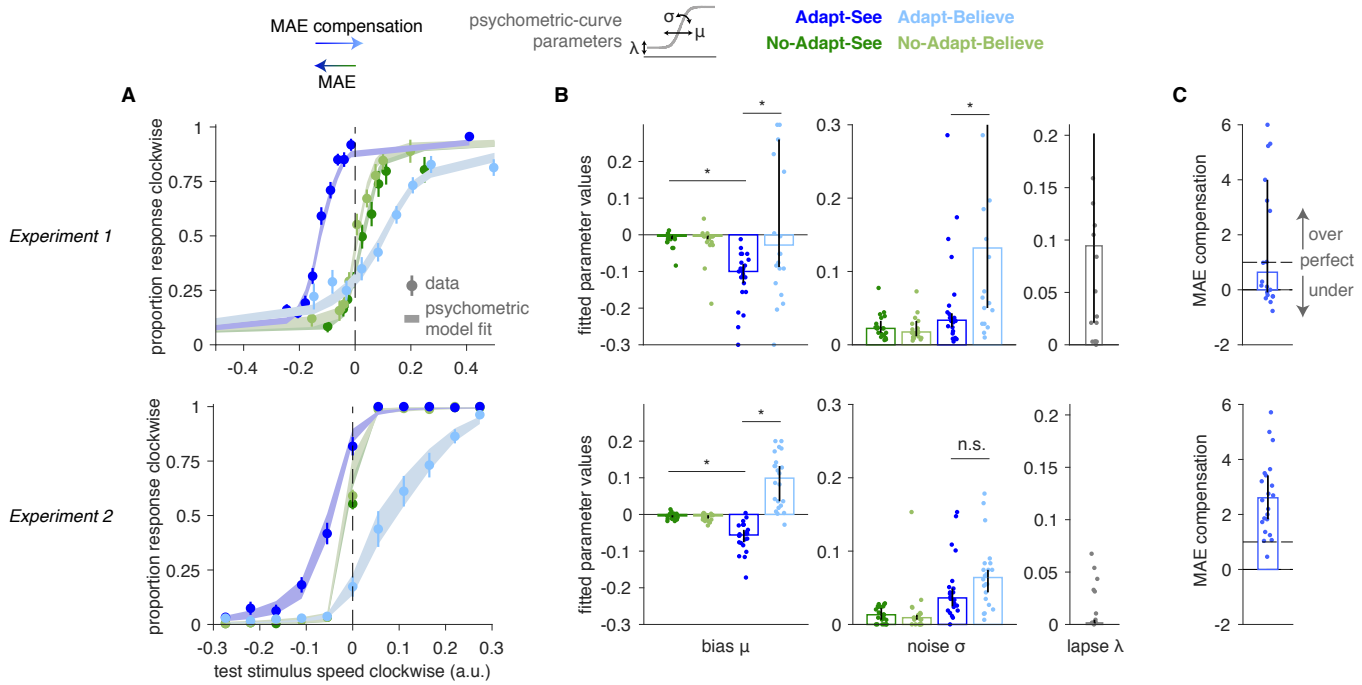

Figure 3: **Changes in psychometric curves demonstrate MAE compensation.** Data shown from Experiment 1 (top) and Experiment 2 (bottom). **A)** Psychometric curves showing mean data ($\pm$ SEM) and averaged psychometric-curve fits. **B)** Fitted psychometric-curve parameter values (bars and error bars representing median $\pm$ bootstrapped 95% confidence intervals [CI]), showing a negative μ bias for Adapt-See (MAE) and μ close to 0 or positive for Adapt-Believe (indicating MAE compensation). * indicates $p < 0.01$. **C)** MAE compensation index normalized by $|\mu_{\text{Adapt-See}}|$ (median $\pm$ CI).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MAE compensation is consistent with an intermediate inferential process

We then tested whether MAE compensation was consistent with perceptual insight. Perceptual insight would imply an adjustment at the intermediate stage of inference that results in the computation of the perceptual-decision variable d (the log-posterior ratio) under Adapt-Believe (Fig. 1C, right). In contrast, late compensation would only imply a change at the late-response stage, such as an offset in k_{choice} (Fig. 1C, center), similar to that induced by response priming or asymmetric reward payouts [24, 25, 32, 33]. The observed shifts in psychometric curves (Fig. 3) are theoretically consistent with either of these two scenarios. However, among these two, only adjustments at the intermediate stage should cause commensurate shifts in perceptual-decision uncertainty - $|d|$ - the negative distance of d from the point of maximal perceptual uncertainty ($d = 0$) - that should be reflected in confidence and RT curves shifting in tandem with psychometric curves [24, 25]. Under late compensation, d should instead be identical in Adapt-See and Adapt-Believe, and so should be any measures reflecting perceptual-decision uncertainty (i.e., confidence and RT); late compensation should thus lead to an isolated shift in the psychometric curves between these conditions (Fig 1C, center) [24, 25, 32].

Consistent with previous work, MAE-related shifts in the psychometric curves were accompanied by commensurate changes in confidence and RT curves (Fig. 4). Furthermore, biases under Adapt-See in these three different measures correlated across individuals (all $0.43 < \rho < 0.87$, all $p < 0.05$; Supplementary Fig. S2, Left). Critically, psychometric curves defining MAE compensation also shifted in tandem with confidence and RT curves (Fig. 4). Under Adapt-Believe, individuals with larger biases in psychometric curves also exhibited larger biases in confidence and RT curves (all $0.63 < \rho < 0.85$, all $p < 0.002$; Supplementary Fig. S2, Right). This shows that MAE compensation involves a shift in perceptual-decision uncertainty, consistent with an adjustment at the intermediate inferential stage and inconsistent with a change restricted to a late response stage.

An alternative explanation for this pattern of compensation is a change at the early stage of sensory encoding, but this seems less tenable. First, participants reported continuing to experience the MAE illusion after receiving instructions and during the Adapt-Believe condition. Second, adaptation induces changes in early sensory neurons, such as reductions in firing rates associated with hyperpolarization [34–38], that are unlikely to be easily reversed at task-relevant timescales.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

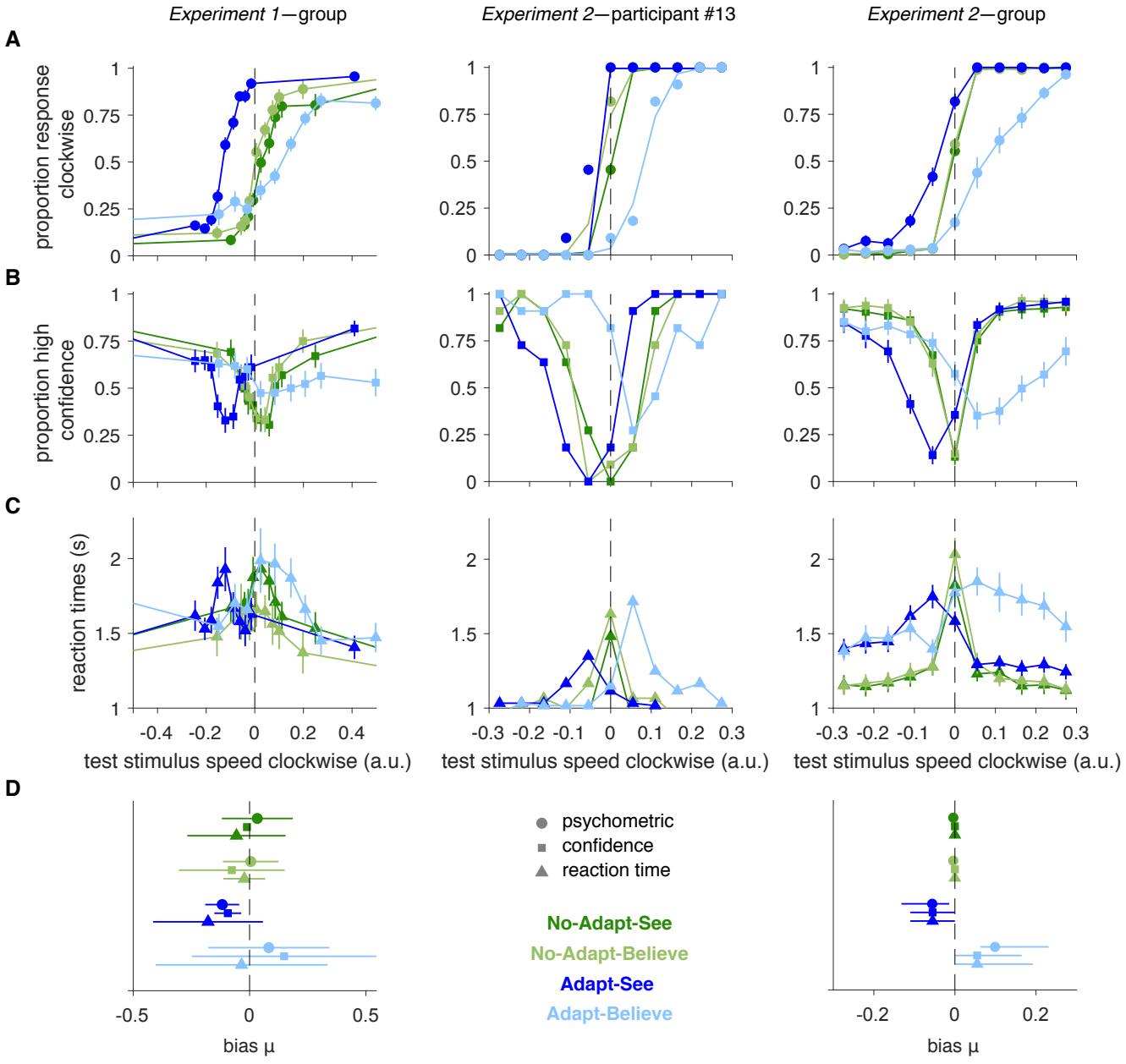

Figure 4: **Psychometric curves for MAE and MAE compensation shift in tandem with confidence and RT curves**. **A)** Psychometric curves. **B)** Confidence curves. **C)** RT curves. Shown are group (mean ± SEM) data for Experiment 1 (left), sample participant from Experiment 2 (center), and group data for Experiment 2 (right). **D)** Concordance of μ bias estimates (median ± CI) across psychometric, confidence, and RT curves for each condition. See Fig. S2 for interindividual correlations.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Modeling supports perceptual insight

To more directly test the model of perceptual insight (Fig. 1, right), we fitted an adapted version of this model to data from Experiment 2 (Fig. 5). We specified the perceptual-insight and alternative models building from an extended Bayesian model of perceptual inference shown to capture variability in confidence reports [16,17] (Methods S2), the architecture of which (Fig. 5A) roughly maps onto the three stages of processing discussed above. An early stage in all models consisted of encoding the stimulus s into a noisy internal representation x, which was offset under the Adapt conditions. We assumed x was equally offset under Adapt-See and Adapt-Believe reflecting an unavoidable effect of the adaptor at this early level. At the intermediate stage, the posterior probability was computed and converted into a perceptual-decision variable d consisting of the log-posterior ratio, which was then used at the late stage to produce a binary choice $\hat{C}$ by comparing d to the threshold k_{choice}. Confidence responses q reflected the posterior probability of the chosen response $p(C = \hat{C}|x)$ [15, 16], binarized via a comparison with the threshold $k_{\text{confidence}}$. Critically, unique to the perceptual-insight model was the incorporation of knowledge about factor A, here the adaptor, in its intermediate inferential stage $P(x|C, A)$ via a shift in the likelihood term $p(x|s, A)$ (Methods S2.2), which produced a shift in d sufficient to compensate for the distortion without need for biasing its response rule ($k_{\text{choice}} = 0$); in contrast, the late-compensation model inferred C without knowledge about A, as $P(C|x)$, requiring a biased response rule ($k_{\text{choice}} \neq 0$) to compensate for the MAE. Because the former scenario produces shifts of psychometric curves and confidence curves in tandem and the latter produces isolated shifts in psychometric curves (Fig. 1D), we deemed this basic model architecture flexible enough to capture the range of relevant behavioral patterns.

Formal comparison of models fitted jointly to choice and confidence reports (Methods S2.4) selected the perceptual-insight model over the late-compensation model, a hybrid model, and alternative models that allowed changes in other inferential variables, including the category prior (Fig. 5A, Methods S2.3). All models allowed for shifts in the internal representation x during the Adapt conditions, relative to the control No-Adapt conditions, via two fixed parameters (one μ_{encoding} value for Adapt and one for No-Adapt), and additionally had condition-specific free parameters for sensory or encoding noise (σ_{encoding}) and confidence thresholds $k_{\text{confidence}}$ (Table S1; Methods S2). Similarly to the perceptual-insight ($\mu_{\text{likelihood}}$) model, the category prior model can compensate for the MAE via an intermediate inferential process manifesting as shifts in tandem, but achieves this compensation through a distinct mechanism; while the perceptual-insight model compensates by flexibly modifying the perceptual-variable d in a way that scales with encoding noise (Equation 16 and Supplementary Fig. S4 (A vs B)), the prior model simply assumes a change in the category prior that produces fixed shifts in d regardless of encoding noise.

Importantly, the winning perceptual-insight model ("$\mu_{\text{likelihood}}$") yielded meaningful parameters and satisfactorily captured the pattern of shifts in tandem associated with the MAE and MAE compensation (Fig. 5B, compare visually with the response bias model fits in Supplementary Fig. S5), accomplishing the latter through condition-specific free $\mu_{\text{likelihood}}$ parameters that shifted the likelihood of x under Adapt-Believe relative to the other conditions (Fig. 5C). Fitted $\mu_{\text{likelihood}}$ differed between Adapt-See and Adapt-Believe ($z = 4.11, p = 4.01 \cdot 10^{-5}$, Wilcoxon sign-rank test), as did the fitted σ_{encoding} ($z = -3.65, p = 2.59 \cdot 10^{-4}$) and $k_{\text{confidence}}$ ($z = 2.58, p = 0.0099$). These parameter differences were meaningful as parameter recovery was successful (Supplementary Fig. S6). While our simplified insight model (Section S2.2) assumed no additional uncertainty related to the Adaptor ($\sigma_A \approx 0$), it captured increased encoding noise in Adapt-Believe relative to Adapt-See via condition-specific σ_{encoding}, likely reflecting a contribution of σ_A. Critically, only the change in $\mu_{\text{likelihood}}$, but not in the other parameters, correlated with psychometric-curve shifts between Adapt-See and Adapt-Believe ($\rho = -0.93, p = 1.25 \cdot 10^{-10}$). Altogether, these results suggest that participants can effectively compensate for an unavoidable perceptual distortion originating at an early sensory stage by incorporating

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

knowledge about this internal distortion at an intermediate stage of perceptual inference. This supports that the observed MAE compensation reflects genuine perceptual insight rather than alternative, or additional adjustments at the late response stage.

To further confirm that our results did not reflect a generic compensation pattern, following previous work [32] we performed a pilot experiment (Methods S1.4) where we replaced the Believe conditions with 'Bias' blocks in which participants were instructed to respond left when uncertain. This response-bias manipulation induced the expected shift in the psychometric curves in the absence of corresponding shifts in the confidence and RT curves, a pattern that was explained by a change in the response-bias parameter k_{choice}, and not by $\mu_{\text{likelihood}}$ (Section S3.7, Supplementary Fig. S7C). These data thus support the notion that perceptual insight can be dissociated from response bias and speak against a generic compensation pattern.

To further validate our winning model of perceptual insight, we examined its ability to capture RTs not used for model fitting (Fig. 5D). We assumed that RTs reflected decision difficulty based on the perceptual-decision uncertainty $-|d|$ derived from the model. We thus tested whether ranked $|d|$ could predict ranked RT on a trial-by-trial basis using a generalized linear mixed-effect model (GLME) controlling for condition (Methods S1.5.2). Indeed, greater uncertainty, smaller $|d|$, correlated with longer RT ($t_{10643} = -10.07, p = 9.49 \cdot 10^{-24}$), providing a good explanation for the RT curves across conditions despite the absence of significant condition effects (all $p > 0.1$; Fig. 5D). Furthermore, this GLME provided a better fit (AIC = 191610, BIC = 191770) than one predicting ranked RT from the ranked objective stimulus strength $|s|$ ($t_{10643} = -8.92, p = 5.08 \cdot 10^{-19}$; AIC = 193390, BIC = 193540) and condition, which in contrast did show condition effects ($p < 0.002$ for Adapt-See and Adapt-Believe). Therefore, RTs could be parsimoniously explained by decision uncertainty reflecting uncertainty of posterior beliefs in the perceptual-insight model.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

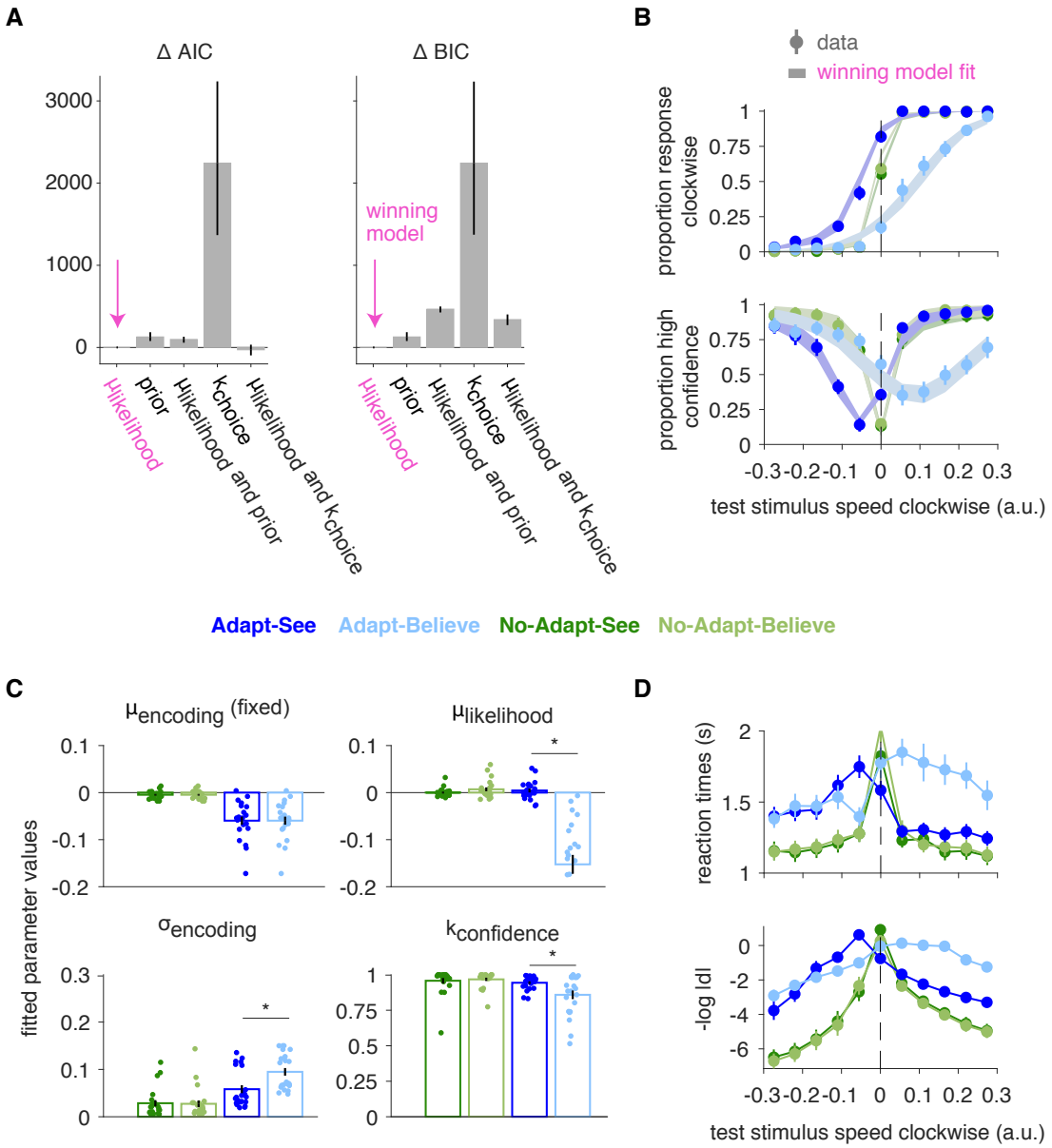

Figure 5: **Modeling of Experiment 2 data supports the perceptual-inference model. A)** Model comparison selects the perceptual-insight "$\mu_{\text{likelihood}}$" model as winning model. **B)** Fits from winning model (shaded areas) capture well choice and confidence curves (mean $\pm$ SEM, as in Fig. 4); corresponding fitted parameter values (median $\pm$ CI) in **C)**. * indicates $p < 0.01$. **D)** RT curves mirror the perceptual-decision uncertainty $-|d|$ from the winning model (shown in log space). Data are mean $\pm$ SEM.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pupillometry further validates the perceptual-insight model

Arousal-related pupil dilation tracks with perceptual-decision uncertainty and relies on norepinephrine and brainstem circuits distinct from motor execution of button presses [26, 39–45], and thus from RTs in this task. We thus used pupillometry as an objective physiological readout of intermediate processing stages (as opposed to late-stage processes involved in planning and executing the button-press response) to corroborate our interpretation of the data in terms of perceptual insight. We specifically hypothesized that pupil dilation in the relevant task periods would mirror RTs and more directly reflect the internal perceptual-decision uncertainty d variable from the perceptual-insight model.

Indeed, pupil-dilation patterns roughly mirrored the confidence and RT curves, showing subtly yet visibly shifted pupil-dilation peaks for Adapt-See versus Adapt-Believe (Fig. 6B) - a remarkable difference considering that these conditions were matched on the adaptor and stimuli and only differed in the required response. In parallel to RT analyses, we ran a GLME predicting ranked pupil area on a trial-by-trial basis as a function of ranked $|d|$ and condition during a decision-related period where the sluggish pupil responses should be most apparent (2000-2500 ms after stimulus onset; Fig. 6A). As expected, smaller $|d|$ indicating greater uncertainty predicted greater pupil area ($t_{10643} = -7.63, p = 2.52 \cdot 10^{-14}$). To control for response contamination, we also performed analyses of response-locked changes in pupil area during a decision-related period (500-1000 ms post response), which confirmed the effect of $|d|$ ($t_{10643} = -5.38, p = 7.28 \cdot 10^{-18}$), even when controlling for RT ($t_{10642} = -2.69, p = 7.11 \cdot 10^{-3}$). Generally, $|d|$ was more closely related to pupil area than the objective stimulus strength $|s|$ ($t_{10643} = -3.83, p = 1.26 \cdot 10^{-4}$), with GLMEs featuring $|s|$ showing poorer fits (AIC $= 197190$, BIC = 197340) and larger condition effects than GLMEs featuring $|d|$ (AIC $= 196950$, BIC = 197100). Finally, these results were confirmed in moving-window GLMEs (Methods S1.5.2) showing strong effects of $|d|$ at the relevant time periods. Response-locked analyses confined to the matched Adapt conditions (Adapt-See and Adapt-Believe) showed consistent results for $|d|$ ($t_{5321} = -6.07, p = 1.34 \cdot 10^{-9}$) that held even when controlling for RT ($t_{5320} = -2.88, p = 0.0039$), demonstrating key model-predicted changes in the pupillometric signature of d that were distinct from RT changes. Leveraging pupil dilation to index a subjective perceptual-decision variable [26, 44], we thus confirmed the predicted changes in the pupillometric signature of this internal variable between Adapt-See and Adapt-Believe, further supporting the validity of the perceptual-insight model.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

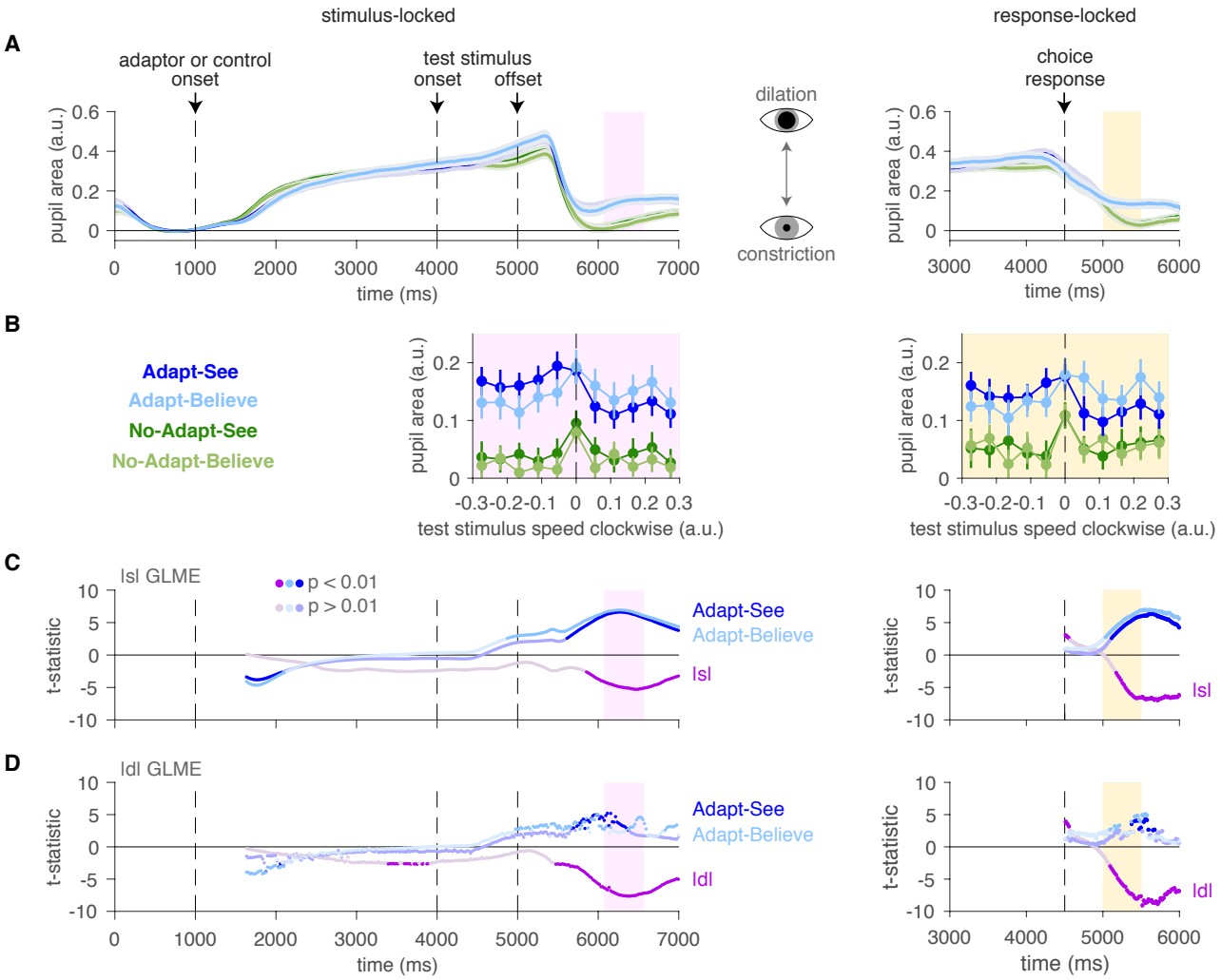

Figure 6: **Pupillometric signature of perceptual-decision uncertainty supports the perceptual-insight model. A)** Time series of normalized pupil area (mean ± SEM) across trials and subjects by task condition. Left: stimulus-locked data (left) showing a decision-related window (pink). Right: response-locked data showing a decision-related window (yellow). **B)** Pupil dilation peaks (mean ± SEM) as a function of stimulus strength show subtle yet visible shifts between Adapt-See and Adapt-Believe (mirroring confidence and RT data in Fig. 4). **C) - D)** Moving-window GLMEs showing the fixed-effect t-statistics time series (see Methods S1.5.2). Brighter colors indicate statistical significance at $p < 0.01$. **C)** The effect of absolute objective stimulus strength $|s|$ is apparent around the decision-related periods. **D)** The effect of model-derived perceptual-decision uncertainty $|d|$ (from the winning model) is apparent around the decision-related periods, and is stronger than that for $|s|$ in **C)**. Effects for Adapt-See and Adapt-Believe (relative to No-Adapt-See) are weaker in **D)** than in **C)**.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

We have presented a formal modeling framework that portrays perceptual insight as the ability to acknowledge distortions in one's internal representations, allowing online adjustments in support of adaptive decision-making. We further designed a controlled "cognitive psychophysics" paradigm [31] this form of in-the-moment reality testing while minimizing memory demands. We first demonstrated that people are generally capable of compensating for distorted percepts associated with a complex MAE illusion. We then leveraged confidence reports and RTs [24, 25] to show that MAE compensation involved shifts in decision uncertainty more consistent with adjustments at an inferential stage. Model comparison and validation lent further support for a model of perceptual insight involving adjustments at this inferential level. The perceptual-decision variable capturing posterior beliefs in our model provided a parsimonious explanation for RT and pupil-dilation patterns, further supporting our conceptualization of perceptual insight and demonstrating humans' ability to deploy such insightful strategies to circumvent internal biases.

As the adage "seeing is believing" implies, conscious perceptual experiences generally dictate one's beliefs. In contrast, our results exemplify a rare case of a compensatory strategy that systematically prevents conscious perceptual experience - of the MAE - from driving beliefs about the true state of the world [46–48]. Put simply, they show that seeing need not imply believing. Under our operationalization, perceptual insight is at the center of this distinction: perceptual insight allows for a flexible mapping between seeing and believing because it incorporates introspective knowledge that internal experiences depend upon factors other than external stimuli and therefore do not always represent the external world faithfully. As such, perceptual insight enables reality testing. The novel systematic dissociation between perceptual experience and belief we achieved empirically, and our multi-stage perceptual-insight framework, are conceptually related to previous hierarchical views of conscious perception and metacognition where beliefs about reality arise from higher-level processes [11, 48]. Our perceptual-insight model also provides an explicit computational account of reality testing that represents an alternative to a previous proposal based on generative adversarial algorithms [49, 50]. Future work is warranted to evaluate and directly compare our framework to these others. More generally, our work suggests that basic principles of perceptual decision-making - which have been successfully applied to explain how individuals discount unreliable external stimuli (a corrupted radio signal or a deepfake video) and deploy metacognitive strategies to evaluate their own performance [51] - can further be applied to elucidate complex forms of introspection that have been elusive in the past.

The assumptions of our perceptual-insight model are broadly consistent with previous theoretical and empirical neuroscience work. As in previous work, the assumption that the MAE predominantly affects the early stage of sensory encoding is based on the observed decreases in neuronal firing rates [37, 52] and fMRI signals in motion-selective sensory regions under MAE [53]. Further, microstimulation of these sensory regions produces shifts of psychometric, confidence, and RT curves in tandem [54, 55] similar to those we and others [24, 25] observed during the MAE. We also assumed that the offset in the sensory measurement under MAE could not be (fully) corrected, since adaptation produces changes in the membrane potential of neurons [34–38] unlikely to be readily malleable. Moreover, the MAE illusion was experienced regardless of explicit knowledge about its illusory nature and of the required (See or Believe) responses. Given this, and consistent with recent work [56], we interpreted the observed shifts in tandem associated with MAE compensation as changes at an intermediate perceptual-decision stage. Consistent with this interpretation, microstimulation of decision regions such as the lateral intraparietal area (LIP) also produces shifts in tandem [57]. An explanation in terms of late compensation seemed less tenable given that changes in the response rule at late stages - likely implemented in distinct downstream regions [58] - can manifest as isolated shifts in psychometric curves [32], (as reproduced in

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

our response-bias pilot experiment; Supplement S3.7), in contrast to our main results. Beyond pointing to an intermediate-stage process, our results supported a model whereby perceptual insight reflects the incorporation of knowledge that factors other than external stimuli (here, the adaptor) influence internal representations. Modeling suggested this knowledge was incorporated at the level of the likelihood term in the inferential process, or equivalently at the level of decoding of the distorted internal representation. Although speculative, this process could be neurally implemented in decision regions (e.g., LIP) akin to optimal-decoding solutions based on synaptic reweighting of sensory representations [59–64], perhaps via neuromodulation-dependent plasticity. Generally, our results extend previous work demonstrating that Bayesian principles for integration of external information govern decision-making [18–21] by showing that similar principles may apply to the incorporation of introspective knowledge of internal distortions.

Potential limitations worth mentioning include our reliance on the dissociation between psychometric-curve shifts in tandem with confidence and RT curves versus isolated shifts in psychometric curves to respectively arbitrate between intermediate inferential biases and late-response biases. While previous work has shown this type of dissociation [24, 25, 32, 65] in line with our pilot experiment (Supplementary Fig. S7), other results are less clear in this respect [33]. Nonetheless, our interpretation of MAE compensation in terms of an inferential process (of incorporating introspective knowledge about the internal distortion) is further supported by evidence (1) that our participants had detailed knowledge about the MAE illusion (its direction and approximate strength; Supplementary Fig. S1); (2) from our model comparison favoring the perceptual-insight model even over other inferential models (e.g., category prior model; Fig. 5A); and (3) from the pupillometric results that more likely reflect internal inferential processes than response-related processes, and which held after controlling for RTs (more likely reflecting the latter). In experiment 2 (although less clearly in experiment 1) participants overcompensated for the MAE, a result we did not necessarily predict. One intriguing potential explanation is that participants expected an MAE illusion stronger than the one they actually experienced during the task (i.e., an inaccurate representation of the prior distribution of A), which could be due to our pre-task instruction using longer-lasting adaptors to better illustrate the MAE (Methods S1.2, S1.3). This possibility would fit comfortably with the perceptual-insight model, but testing it will require future work systematically manipulating expectations about MAE strength. Furthermore, the interindividual variability in the MAE compensation hints at a variety of factors – some explicitly identified in our perceptual-insight model – potentially driving overuse or underuse of introspective knowledge in health and illness.

A key future goal of this research is to advance our understanding of elusive insight impairments that are central to psychosis [1, 2, 4]. In this context, our model may suggest that impaired insight in patients with psychosis stems from a failure to recognize that, in contrast with lifelong experience, their internal representations do not match the external world. This failure could thus relate to the sort of inflexible belief-updating proposed to underlie other aspects of psychosis [66–68]. Overall, this work will hopefully provide a foundation for the quantitative study of insight impairments and contribute to the development of objective markers of subjective experience, e.g., based on pupillometry [69, 70].

In conclusion, we have developed a modeling framework and a controlled psychophysics paradigm that captures perceptual insight as a quantifiable compensation for distorted perception. Our results collectively suggest that perceptual insight can be used to counteract upstream distortions in sensory measurements via downstream adjustments at an inferential readout stage. By decoupling percepts and beliefs, perceptual insight reveals a key interface that may prove helpful in advancing our understanding of introspection and conscious perception.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S1 Methods

S1.1 Experimental approach

Experiments 1 and 2 used variants of our perceptual-insight task to quantify compensation and test our candidate mechanism of perceptual insight. We will first present the core components of our perceptual-insight task based on the motion after-effect (MAE) and discuss small differences across experiments, first in brief and then in more detail.

S1.1.1 Perceptual-insight task

The task, trial structure and overall experiment structure were mostly shared across the two experiments. The exact task design of Experiment 2 is in Fig. 2. The task requires the participants to perform left-right motion discrimination: is the test spiral rotating counterclockwise (left response) or clockwise (right response)? In order to measure implicitly the strength of the motion after-effect, we used the nulling method [29, 71–73], which entails the presentation of two spirals, the first one being the adaptor, and the second one being the test stimulus. Observers are then required to give their responses about the second test spiral. Stimuli are full screen spiral textures, adapted from [74], with a small circle marker in the center to indicate fixation. As depicted in Fig. 2, on every trial, participants first saw a fixation screen for 1000 ms, followed by a screen of 3000 ms with the first adaptor spiral (moving with a fixed high speed on Adapt blocks or static in No-Adapt blocks), and then a screen for a maximum of 1000 ms with the second test spiral (rotating at variable speed). Afterwards, a first change in the color of the central marker prompted the left/right choice about motion direction, and then another change in its color prompted a confidence report about their choice (very sure [high confidence] or not very sure [low confidence]) (Fig. 2). Overall, the experimental task had a 2-by-2 design, with two types of conditions No-Adapt (static first spiral) and Adapt (rotating first spiral) - and 2 types of responses required, which direction participants See the second spiral moving or which direction they Believe the second spiral is actually moving. The behavioral measures of interest shared across both experiments were: choice responses (left/right), confidence responses (high/low) and reaction times.

Overview of experiments 1 and 2.

Experiments 1 and 2 were in-person and both measured behavior; Experiment 2 additionally employed eye-tracking. Experiment 2 had three main developments relative to Experiment 1. First, it made use of enhanced instructions which contained more informative feedback. Second, it presented the participants with a fixed set of stimuli drawn from a uniform distribution (with bounds informed by the data from Experiment 1), thus making the data amenable to fitting the proposed Bayesian model of insight. Third, by using eye-tracking, we ensured that participants maintained fixation during the trial and also measured their eye positions and pupil sizes continuously throughout the experiment. Specifically, during the fixation, adaptor and test stimulus periods (a total of ≈ 5 s), if the participants either blinked or moved their eyes relative to the screen away from a circle centered at fixation with radius 3 degrees of visual angle (dva), the trial stopped and the same trial was again presented immediately after until the participant completed it without blinking or breaking fixation. During a 3 s post-stimulus period, if the participants either blinked or moved their eyes away from a circle centered at fixation with radius 24 dva, the trial was also stopped and the same trial was presented again. This approach served to control for the potential influence of eye movements and blinks [75] on the motion after-effect and to rule out certain strategies that could be used to minimize it.

Other minor differences across experiments 1 and 2 were: experiment 1 had a slightly shorter fixation window (890 ms vs 1000 ms) and different colors of fixation circles and choice prompts. Additionally,

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the adaptor before the blocks was presented for 15 s in experiment 1 and 30 s in experiment 2. 5 out of 22 participants from experiment 1 were presented with the adaptor spiral rotating clockwise; their data was combined with data from the other participants by flipping directions accordingly (such that the reference direction of the adaptor was counterclockwise). Note that all the reported effects were present even if we exclude these 5 participants from experiment 1.

S1.2 Experiment 1

Participants.

25 participants were recruited. 3 participants were excluded after the instructions because they did not experience the illusion as their responses in the second part of the training were not consistent with the illusion on a sufficient number of trials (less than 8 out of 10). The remaining 22 participants that completed the task and were used for analyses comprised 11 males, and 11 females, and had a median age of 26 years old (range 21 to 50). All participants provided informed consent. The study conformed to the Declaration of Helsinki and was approved by the Institutional Review Board of New York State Psychiatric Institute.

Apparatus.

The stimuli were displayed on a 13-inch MacBook Pro laptop (2017 model) in a dark psychophysics room. The width of the viewable portion of the screen was 11.97 inches (30.4 cm) and the screen resolution was 2560 $\times$ 1600 pixels and 60 Hz refresh rate (1 frame lasting approximately 16.67 ms). The MacBook Pro laptop had installed Matlab 9.6 (2019a, MathWorks, Massachusetts, USA) with the Psychtoolbox extension, version 3.0.15 [76–78]. Matlab and Psychtoolbox controlled the presentation of the stimuli and recorded the participants' responses. Participants were seated at a distance from the laptop of approximately 20 cm.

Instructions, demonstration, training.

This section contained instructions and training trials and was divided into 8 parts with a few additional quiz questions. Collectively, these elements were used to gradually build and ensure an adequate construal of the task. In the first part, participants were asked to perform 10 trials of left/right motion discrimination of a single rotating spiral and received feedback after each trial. Each participant was asked to perform this part until performance was higher or equal than 70% correct, for a maximum of 3 times. The second part entailed 10 trials showing two sequential spirals, the first one rotating for 3 s and the second one static for 1 s, and participants were asked to report whether they saw the second spiral moving left or right. Here they received feedback about whether the responses were consistent with the motion-after effect illusion. The participants were asked to perform this part until performance was consistent with the illusion in 80% of the trials, for a maximum of 3 repetitions. The third part repeated the structure of the second part, but importantly entailed the presentation of a sound to mark the onset of the second spiral. Participants were encouraged to convince themselves that the second spiral was indeed static and that looking at the first moving spiral caused them to perceive the second static spiral as moving in the opposite direction. The fourth part built on the third part by introducing motion in the second spiral and asking the participants to report across 10 trials whether they saw the second spiral as moving left/right. Participants were told not to be concerned with the actual motion in the second spiral, foreshadowing that this would be relevant to the 'Believe' condition introduced next. The fifth part introduced the Believe condition. Across 10 trials, participants reported which direction they believed the second spiral was actually moving. To ensure they had explicit knowledge about the illusion, participants were then asked to respond to the following questions: "Imagine that the first spiral was moving [clockwise/counterclockwise], and then you saw the second spiral as still. Which way do you believe the second spiral was actually moving?". The sixth part introduced the confidence responses ("very sure"/"not very

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sure") to augment the left/right responses when discerning the direction of motion of a spiral across 10 trials. Participants were told on how many trials they responded "not very sure" so they could reflect on their confidence responses and attempt to balance them, versus disproportionately responding with just one button (i.e. 9 vs 1). The seventh part again presented two spirals and asked participants across 5 trials to report the direction they saw the second spiral moving and then their confidence about this response. The eight part finally asked participants across 5 trials to report the direction they believed the second spiral was moving and their confidence about this response.

Payment.

In Experiment 1 participants received $30 for the first hour and $20 for each additional hour. They were instructed to try their best in all conditions and incentivized with a bonus of $10 determined based on adequate compliance with task instructions and the experimenter's examination of the data, including consistency in behavior.

Stimuli.

Adaptors of complex motion (e.g., rotational) have been shown to induce stronger adaptation effects relative to ones of simple motion (e.g., translational) [27, 28]. We aimed to choose spiral stimuli to achieve the strongest motion after-effect (or spiral after-effect) illusion with the shortest presentation times of adaptor and test spirals. Within spiral stimuli, we considered their type (Archimedean [79] vs logarithmic [80, 81]), spatial frequency and temporal frequency [82, 83]. Based on the literature and pilot data, we decided to use Archimedean spirals, with spatial frequency parameters $w_1 = 30$ and $w_2 = 3$. A w_1 value of 30 corresponds to 30 pixels being traversed during a cycle. Based on w_1, screen width, the distance from the screen and the screen resolution, we calculated the spatial frequency to be 1.16 cycles/°.

The Archimedean spiral equation is:

$$\text{spiral} = \frac{\text{white}\left(1 - \cos\left(\frac{r}{w_1} + \theta\, w_2\right)\right)}{2} \tag{1}$$

A third parameter captured the temporal frequency of the motion of the spiral, was set to $w_3 = 9$ for the adaptor and varied for test stimuli (Experiment 1: [-2,2] and Experiment 2: [-0.3,0.3]), thus determining the velocity or speed of spirals, as the spatial frequency was always the same across trials (see Equation 2). To implement rotational motion, the spiral was offset relative to the previous screen by a particular angle value, determined by w_3. Given the refresh rate of 60Hz, the temporal frequency was 1.5 cycles / s. Velocity (here used interchangeably with speed) of the adaptor spiral, given by the ratio of the temporal frequency and the spatial frequency, was 1.293 °/s.

$$\text{velocity } (^\circ/\text{ s}) = \frac{\text{temporal frequency (cycles / s)}}{\text{spatial frequency (cycles / }^\circ)} \tag{2}$$

We implemented the adaptor and test spirals based on [74]. In contrast to some other studies (i.e. [84–86]), we decided to keep the contrast of the adaptor and the test spiral both equal to the maximum (1), to avoid introducing interindividual differences due to contrast sensitivity and luminosity.

Test stimulus generation.

The speed of the test stimulus was based on the participant's previous responses according to an adaptive procedure, a type of Bayesian staircase, applied separately for each of the 4 conditions. Each one of the 4 conditions contained 120 trials, divided into 2 blocks of 60 trials each. We used the Matlab implementation [87], based on [88] with extensions to include the lapse rate [89, 90] (this exact staircase was used in [91]). This procedure maintains a posterior distribution over the parameters and updates it after each trial. The next stimulus value is chosen to minimize the entropy of the updated posterior given the stimulus, averaged over the participant's possible responses weighted by their respective probabilities

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[88]. We defined the space of parameters for the computation of the posterior distribution in [87] as follows: for μ, we used a linear grid of 51 points from -0.5 to 0.5, for σ a logarithmic grid of 25 points from 0.001 to 0.5 and for λ a linear grid of 25 points from 0 to 0.3.

The test stimuli were substantially slower than the adaptor. Each one of the 4 staircases generated on every trial a value w_3 within the range $[-2, 2]$ that dictated the temporal frequency and the velocity (speed) of the test stimulus. The test stimulus values thus corresponded to velocities in the range [-0.285 °/s, 0.285 °/s]. The actual values selected through the staircase and used in the experiment had a narrower range (typically within $[-0.5, 0.5]$ in a.u.; see Fig. 3A, top). These values amount to an adaptor/ test velocity ratio of 4.5 or higher, consistent to some extent with values in the range [2, 5], which were shown to achieve the maximum relative bias in previous work [60].

Trial structure.

Trials had the basic structure described in Fig. 2. Participants first saw a fixation screen of mid-level gray (RGB: [128 128 128]) for 890 ms with a central small white circle. This was followed by a screen of 3000 ms with the first adaptor or control spiral (moving with a fixed high speed or static), which was followed by a screen lasting a maximum of 1000 ms (or until the participant responded if the reaction time was under 1000 ms) with the second test spiral (moving with variable speed as determined by the adaptive staircase procedure). The transition from the first (adaptor or control) spiral to the second spiral (test stimulus) was marked by the change in color of the small fixation circle from white to either yellow, for the See conditions, or blue, for the Believe conditions. See and Believe trials were blocked (see below), but this color coding provided an additional reminder of the required responses. If the participant submitted a left/right choice response during the 1000 ms when the test stimulus was on the screen, the next screen was immediately presented featuring a pink-colored fixation circle prompting for a confidence response (requiring an up/down key press). If the participant did not respond within 1000 ms, the test stimulus screen was replaced with a gray screen with no changes in the fixation-dot color (yellow or blue depending on the trial type) which was displayed until response. There was no deadline, making the task self-paced. After response, the screen with the pink-colored fixation circle prompted for the confidence response until this was submitted. In Experiment 1, if the total post-stimulus time (including the left/right choice and up/down confidence response times) was shorter than 2000 ms, the fixation screen was presented to ensure a minimum post-stimulus total time of 2000 ms.

Experiment structure and block order.

Before the experimental task trials, participants performed a short illusion-reproduction task. They saw a 15 s rotating spiral followed by a second static spiral and were asked to remember the strength of the illusory motion on the static spiral. Afterwards, they were asked to reproduce the strength of the illusion in 10 consecutive trials using the computer trackpad. During the reproduction, they saw the static spiral with a green fixation circle and controlled its speed using the trackpad (scrolling away from the center to control direction and speed) until the speed of the static spiral matched that of the illusory motion they had just experienced. These illusion-reproduction estimates were collected once again after the Adapt blocks, and together they served as confirmation that participants had explicit knowledge about the motion-after effect illusion, including its direction.

The main task had a fixed order and was structured as follows: Adapt blocks were presented first, including two Adapt-See blocks followed by two Adapt-Believe blocks, each starting with an extended 15 s presentation of the adaptor spiral (as in [30, 92]) before the presentation of the experimental trials. These 4 Adapt blocks were followed by 10 more trials of illusion reproduction. Afterwards, participants took a break of at least 10 minutes to minimize any possible lasting effects associated with adaptation. After the break, participants performed the No-Adapt blocks: two No-Adapt-See blocks and then two No-Adapt-Believe blocks, each starting with an extended 15 s presentation of a control static spiral. Adapt blocks were presented before the break to reduce fatigue for the active conditions (although 3 participants

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

performed No-Adapt blocks before the break instead and we did not observe systematic differences between these and the rest other than general speeding of responses as the experiment progressed). The fixed order of blocks also served to remind participants of the block-relevant instructions before starting the experimental trials and required responses and to minimize switching costs.

S1.3 Experiment 2

Participants.

26 participants were recruited. One participant was excluded due to medication, 2 because they did not experience the illusion, and one dropped out without completing all procedures. The 22 completers comprised 6 males and 16 females, and had a median age of 24.5 years (ranging from 18 to 34 years old). All participants provided informed consent. The study conformed to the Declaration of Helsinki and was approved by the Institutional Review Board of New York State Psychiatric Institute.

Apparatus.

The experiment took place in a dark psychophysics room. The computer used was a Mac mini (Late 2014 model) with a 3 GHz Intel Core i7 processor, 16 GB 1600 MHz DDR3 memory and Intel Iris 1536 graphics. The width of the viewable portion of the screen was 59 cm and the screen resolution was 1920 $\times$ 1080 pixels and 60 Hz refresh rate (1 frame lasting 16.67 ms). The Mac mini had installed Matlab 8.3 (2014a, MathWorks, Massachusetts, USA) with the Psychtoolbox 3.0.13 extension [76–78].

In addition to Experiment 1, in Experiment 2 we monitored participants' fixation and recorded their eye movements and pupil sizes. We used a remote infrared video-oculographic system (EyeLink 1000 Plus version 5.15; SR Research, Ltd, Mississauga, Ontario, Canada) with a 1 kHz sampling rate. Participants were seated such that the screen distance (eye to center of the screen) was approximately 66 cm. For the majority of the participants, we set the heuristic filtering option 'ON'. The eye tracker was calibrated using the 9-point calibration routine before every block.

Instructions, demonstration, training.

This section contained the same 8 parts as in Experiment 1 in addition to a new part and other enhancements. The instructions were enhanced (based on our experience and feedback from participants in Experiment 1) and provided more detailed and informative feedback so that participants could have more precise knowledge about the illusion. After the first three parts transpired as in Experiment 1, participants first experienced how they can control the speed and direction of a static spiral by moving the mouse farther away to the right or to the left. In the fourth part, participants practiced 10 trials of illusion-reproduction. This was not included in the training from experiment 1, just in the experiment 1 itself. As in experiment 1, participants saw the first rotating spiral for 30 s, then a static spiral and were told to remember the strength of their illusion. Then they saw 10 static spirals each with a green fixation circle and each time controlled its speed using the mouse until the speed of the static spiral matched that of their illusory motion. This served as training for the illusion-reproduction trials from the experiment itself. In the fifth part, participants saw an adaptor spiral followed by test spiral and performed motion discrimination, as in Experiment 1. Here however they also received feedback with the actual velocity of the test spiral, presented as the percent of the speed of the adaptor spiral, and the direction of motion (same or opposite) relative to the adaptor spiral. In the sixth part introducing the Adapt-Believe condition (corresponding to the fifth part from Experiment 1), participants saw an adaptor spiral followed by a second test spiral with different motion speeds and reported their beliefs about its true motion. Here, participants received feedback on their accuracy based on the true motion of the test spiral and again received information about the actual velocity of the test spiral. Participants next completed the same quiz as in Experiment 1. The last three parts introducing confidence responses were as in Experiment 1, but had more trials (10 vs. 5) and additional feedback of the actual speed of the test spiral relative to

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the adaptor spiral. The eight part consisted of 10 Adapt-Believe practice trials. If response accuracy on these trials was below 70%, participants were asked to repeated this set of 10 trials until their accuracy was above 70 %, for a maximum of 3 attempts. Overall, the instructions and training for Experiment 2 were similar to those in Experiment 1 and mainly differed in that they contained more detailed feedback and information including the actual relative speed of the second test spiral.

Payment.

Participants were compensated with \$30 for the first hour, and \$10 for each additional hour. They were instructed to try to be as consistent as possible in all conditions and as accurate as possible in the Believe conditions. They were incentivized with a bonus of \$30. As in experiment 1, this bonus was determined based on adequate compliance with task instructions and the experimenter's examination of the data, including consistency in behavior.

Stimuli.

The stimuli were also Archimedean spirals with the same parameters as in Experiment 1: $w_1 = 30$ and $w_2 = 3$. The temporal frequency was the same for the adaptor spiral ($w_3 = 9$) but differed for the test spirals in that was drawn from a uniform distribution between $[-0.3, 0.3]$. The screen width, distance from the screen and the screen resolution were different from those in Experiment 1. Given this, the spatial frequency in Experiment 2 was 1.14 cycles/° and the velocity of the adaptor was $1.5/1.14 = 1.31$ °/s.

Test stimulus generation.

In contrast to the adaptive staircase in Experiment 1, Experiment 2 used test stimuli drawn from a uniform distribution (in a pseudorandom order fixed across participants). The bounds of this uniform distribution were informed by the empirical distributions of test stimuli from Experiment 1. The values were between $[-0.3, 0.3]$, corresponding to test stimulus velocities between $[-0.044°/s, 0.044°/s]$. This range was maintained across conditions to facilitate direct comparison of responses and provide a parametric distribution satisfying assumptions of, and allowing for model fitting based on, Bayesian models. Each of the 4 conditions contained 121 trials, divided into 2 blocks of 60 and respectively 61 trials. We chose 121 trials (vs. 120 in Experiment 1) to include the velocity value of 0, corresponding to a static test spiral.

Trial structure.

As in Experiment 1, trials had the basic structure described in Fig. 2. A major difference was that fixation was enforced for a time window encompassing the fixation, adaptor, and test stimulus periods (5 s total) and blinking was also enforced for this time window plus an additional post-stimulus interval of about 2 s, for a total of about 7 s. If participants blinked or deviated their eyes outside a circle of 3 dva centered at the fixation central dot during this period (fixation, adaptor or test stimulus) or blinked or deviated their eyes outside a circle of 24 dva during the post-stimulus period, the trial was immediately interrupted and the same trial was repeated until responses were provided without blinks or breaking fixation. Participants were encouraged to keep their eyes within the fixation area throughout the trial and were encouraged to blink at the end of their trial and briefly rest their eyes as needed. The moment they decided to resume fixation dictated the initiation of the next trial.

Participants first saw a fixation screen of dark gray (RGB: [85 85 85] vs. [128 128 128] in Experiment 1) for 1000 ms (vs. 890 ms in Experiment 1) with a central gray dot of RGB [28.7 28.7 28.7], equivalent to [10 0 0] in CIE-L*ab color space, with L = 10. The darker color of the fixation screen (RGB [85 85 85] equivalent to [36 0 0] in CIE-L*ab, with L = 36) was close in overall luminance to the spiral texture, which minimized spurious changes in pupil size due to luminance and minimized eye-muscle strain under the more demanding fixation-enforcement procedure. Next, a screen of 3000 ms with the first adaptor or control spiral was presented. This was followed by a screen lasting a maximum of 1000 ms with the second test spiral (moving with variable speed). The transition from the first spiral (adaptor or control)

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to the second spiral (test stimulus) was marked by a change in color of the small fixation circle, from gray to a yellow of RGB [81 75 0] and respectively [31 -6 40] in CIE-L*ab, with L = 31, close to the luminance of the gray of the screen (L = 36). This yellow dot signaled the onset of the test spiral in both See and Believe conditions (unlike in Experiment 1, where different colored dots were used for See vs. Believe) to keep stimulation identical between these conditions. In Experiment 2, even if the participant made a left/right response about the test stimulus faster than 1000 ms, the test stimulus was always kept on screen for the entire 1000 ms (precisely 1033 ms due to the refresh rate). If the participant did not respond within this window, the test stimulus screen was replaced with the same gray screen maintaining the yellow fixation circle until response. After the left/ right response was recorded, a new screen with a blue fixation circle (the shade of blue, RGB [0 85 140] and respectively CIE-L*ab [34 0 -36], with L = 34, being approximately isoluminant with the yellow in the previous fixation circle) was presented prompting for a confidence response (requiring an up/down key press). If the total post-stimulus time was shorter than 3000 ms, the fixation screen was presented to ensure a minimum post-stimulus total time of 3000 ms (vs. 2000 ms in Experiment 1). These differences between the experiments ensured that stimulation across conditions was identical between Adapt-See and Adapt-Believe, which allowed direct comparisons of pupil responses (see below).

Experiment structure and block order.

The experiment structure and block order were comparable to Experiment 1. Adapt blocks started with an 30 s presentation of the adaptor spiral and No-Adapt blocks with a 30 s presentation of the control spiral (vs. 15 s for each in Experiment 1). Adapt blocks were completed before No-Adapt blocks. The only difference from Experiment 1 was that the No-Adapt blocks (No-Adapt-See vs. No-Adapt-Believe blocks first or second) were randomized across participants to minimize fatigue, rushing, and task familiarity effects in the control conditions. Adapt-See were followed by Adapt-Believe blocks in a fixed (non-randomized) order because the Adapt-Believe condition involved all elements in Adapt-See plus maintaining and using information about the illusion, and thus benefited from additional exposure to the basic task setup in Adapt-See.

The fixed order of some of our conditions could in principle have led to differential fatigue or practice effects, and systematic differences in RTs, across conditions, but the results of RT analyses speak against this possibility (condition effects on RT were not significant in the GLME featuring $|d|$).

S1.3.1 Task and illusion comprehension checks

Several checks were used to ensure that participants understood the task and had explicit knowledge about the illusion. First, participants' verbal expression of their understanding of the task and illusion in their own words was deemed to be acceptable by the experimenters. Second, their responses on the second part of the training had to be consistent with the illusion in at least 80 % of the trials. If this did not happen within 3 repetitions, participants were paid for their time so far and excluded. Third, responses on key questions of the quiz during training were consistent with their proper understanding of the illusion or prompted clarification until understanding was established by the experimenter. Fourth, the training from Experiment 2 included a performance criterion of 70% accuracy during Adapt-Believe practice trials that demonstrated directional knowledge of the illusion. Fifth and most definitively, in both experiments participants demonstrated explicit directional knowledge of the illusion in illusion-reproduction trials: the estimated strength of their motion-after-effect illusion showed that they expected to experience illusory motion opposite to the direction of the adaptor, reasonable in magnitude, and consistent within subject (Fig. S1).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S1.4 Control Experiment

To test whether observers could potentially employ a compensation strategy distinguishable from the one prescribed by the perceptual-insight model, we performed a pilot control experiment. We specifically used a response-bias manipulation combined with MAE to test whether we could observe a compensation-like pattern in the form of isolated shifts in the psychometric function in line with the late-compensation "k_{choice}" model (Fig.1, Fig.5A, Table S1) or instead whether the response-bias manipulation would result in similar behavioral patterns (shifts in tandem) to those in experiment 2. To this end, we used a response-bias manipulation based on previous work (experiment 1 in [32]) in addition to our adaptation manipulation. Specifically, observers performed the No-Adapt-See and Adapt-See conditions as in experiment 2 and additionally performed No-Adapt-Bias and Adapt-Bias conditions (instead of the Believe conditions) were they were instructed to choose a left counterclockwise response (in the direction that would compensate for the clockwise MAE) when they were uncertain about the direction of the test stimulus, via the instruction: "If unsure, press Left". Instead, during the See conditions in the control experiment (and in all conditions in experiment 2) the instruction was: "If unsure, guess" (aiming to balance guesses across left and right responses).

For this experiment, 11 participants were recruited. 3 participants were excluded due to low quality of their data, specifically their responses not being captured well by psychometric curves. The 8 completers comprised 5 males and 3 females, and had a median age of 25 years (ranging from 21 to 47 years old). Participants were seated at a screen distance of 57 cm (versus in Experiment 2, 66 cm).

S1.5 Statistical analyses

All statistical analyses were performed in Matlab R2019a. Assumptions of parametric tests were violated for most of the main variables, so our main analyses consisted of Wilcoxon rank-sum and signed-rank tests (Matlab's `ranksum` and `signrank`) or Spearman correlations. To visualize the psychometric, confidence, reaction-time and pupil curves as a function of test stimulus speed, we divided the values into 11 bins (with 11 data points each). Data was visualized as mean and standard error of the means (usually across participants) or as median and 95% bootstrapped confidence intervals where non-parametric tests were used. To calculate the 95% bootstrapped confidence intervals, we took samples of the same size as the data with replacement across 5000 iterations using Matlab's `randsample` and calculated the sample median for each iteration. The 95% confidence intervals were based on the 2.5th and 97.5th percentiles of the distribution of medians across iterations.

S1.5.1 Estimation of bias μ from confidence curves and reaction-times curves

The bias μ for the psychometric curves was based on psychometric-curve fitting presented below. We also estimated the bias μ from confidence curves and reaction-times curves for each individual (Fig. 4). As these curves were not well captured by Gaussians in our data, we estimated the bias μ as the mean value of the bin (bin center) with the minimum confidence or maximum RT for each individual. If several bins tied on the minimum or maximum values, the bias μ was calculated as the mean of these bin centers.

S1.5.2 Generalized linear mixed-effect models (GLMEs)

We used GLMEs to quantify the relationships between reaction times and either absolute stimulus strength $|s|$ or absolute perceptual-decision certainty $|d|$ (Fig. 5D), as well as the relationship between pupil area and these independent variables (Fig. 6C,D). The main independent variables ($|s|$ or $|d|$ depending on the GLME) were ranked. All GLMEs (implemented using Matlab's `fitlme`) included fixed effects for

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the intercept, for the independent variables of interest ($|s|$ or $|d|$ depending on the GLME), and for condition (coded as 3 dummy variables for each condition with No-Adapt-See as the reference condition), as well as random intercepts and slopes for all variables. The following GLMEs were used (formulas in Wilkinson notation):

$$\text{Reaction time} \propto 1 + |s| + \text{Condition} + (1 + |s| + \text{Condition}|\text{Subject})$$

$$\text{Reaction time} \propto 1 + |d| + \text{Condition} + (1 + |d| + \text{Condition}|\text{Subject})$$

$$\text{Pupil area} \propto 1 + |s| + \text{Condition} + (1 + |s| + \text{Condition}|\text{Subject})$$

$$\text{Pupil area} \propto 1 + |d| + \text{Condition} + (1 + |d| + \text{Condition}|\text{Subject})$$

$$\text{Pupil area} \propto 1 + \text{RT} + |d| + \text{Condition} + (1 + \text{RT} + |d| + \text{Condition}|\text{Subject})$$

For the GLMEs with pupil area as the dependent variable, we computed the average of the normalized pupil area within a 500 ms time window of interest (respectively pink and yellow regions in Fig. 6A). To check that these results were robust to the choice of window and for intepretability, similar to previous work [93], we used a sliding window of 200 ms which we applied in steps of 20 ms across the entire pupil area time series (Fig. 6C, D).

S1.6 Eye tracking, focus on pupillometry

In Experiment 2, we monitored and recorded eye position and pupil size using an infrared video-oculographic system (Eyelink 1000 Plus; SR Research, Ltd, Mississauga, Ontario, Canada) with 1000 Hz sampling rate. We recorded monocular data. Stimulus presentation and response collection were controlled by a Mac computer running Matlab 7.1 (MathWorks, Massachusetts, USA) with Psychtoolbox 3 [76–78] and EyeLink software [94]. Participants used a chin rest to stabilize their head. The chin rest which was located at approximately 55 cm from the top knob of the desktop mount camera and the distance from the eye to the center of the screen was 66 cm. The eye tracker was calibrated using the 9-point calibration routine before every block.

Fixation monitoring.

Fixation was strictly imposed within each trial during a period encompassing the fixation, adaptor and stimulus periods. If the participant's gaze deviated outside of a circle of radius 3 degree of visual angle (dva) centered at fixation, or if they blinked, the trial with the exact same stimulus value was restarted (see Trial Structure above for exact details). This approach ensured that participants fixated during the adaptor and stimulus presentation and thus maximized adaptation effects. Additionally, if participants responded earlier than 3000 ms after stimulus onset, blinks post response still lead to a restart of the trial. Thus, participants avoided blinking for about 8000 ms in each trial. We instructed participants to blink as much as needed and rest their eyes sufficiently after each trial, and that resuming fixation afterwards would initiate the next trial.

Pupillometry data analysis.

For each participant and block, we defined the relevant task periods within a trial based on the Eyelink timestamps and concatenated the pupil-area time-series across trials. Mean time-series per condition are shown in Fig. 6A, left. Artifact removal was conducted based on [95]. Segments of the time-

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

series within a trial where pupil dilation speeds exceeded a threshold were removed. This threshold was determined based on threshold $= median(\text{dilation speeds}) + n \cdot MAD(\text{dilation speeds})$, where MAD is the median absolute deviation and n is a multiplicative factor which we set to 16. In keeping with recommendations [95], we chose this value since our empirical checks showed that it was sufficient to eliminate visible outliers in our data (note that our strict control for blinks yielded relatively clean time series). Each removed segment exceeding this threshold was replaced with interpolated values based on 20 ms before and 20 ms after the segment. Next, the pupil-area time-series were filtered on each trial using a Butterworth filter of order 2 and bandpassed between 0.01 and 10 Hz, implemented in two steps as in [26] with Matlab's functions `butter` and then `filtfilt`. For normalization, based on previous work [93, 96, 97], we first subtracted the average baseline period - specifically the mean pupil dilation in the last 400 ms of the fixation period - from the pupil-area time-series for each trial. We concatenated all baseline-subtracted time-series across trials and blocks corresponding to each condition. For each participant, we then calculated the maximum value across all conditions and divided the time-series for each condition by that value.

Our main time windows of interest lasted 500 ms (consistent with previous work [41, 93]) and started either at 2000 ms post-stimulus or at 500 ms post-response to capture the delayed peak in pupil responses (approximately 1000 ms) while ensuring pupil stabilization following luminance transitions (e.g., between the brighter stimulus screen and the darker gray response screen).

For the moving window analyses in Fig. 6C and 6D, we applied the GLMEs across sliding windows of 200 ms each, shifted in steps of 20 ms. Note that in the response-locked data from Fig. 6C and 6D, we do not show pre-response t-statistics as they are not interpretable. This is because differences in reaction times create artifactual differences in pupil dilation due to the latency of pupil stabilization following test stimulus offset (Fig. 6A).

S1.7 Psychometric curves

Psychometric curves and parameters. We fit psychometric curves based on Gaussian cumulative density functions (cdfs) to the observers' left/right responses as a function of the test stimulus speed (Fig. 3A). s denotes the test stimulus value on a given trial (in arbitrary units of speed, with negative sign indicating counterclockwise motion and positive sign indicating clockwise motion; conversions to standard units of velocity are provided above). Test stimuli ranged between [-2, 2] in Experiment 1 and [-0.3, 0.3] in Experiment 2. We assumed the Gaussian cdf psychometric curve formula [98], in which a response r depends on a stimulus s and the parameters - $\mu,\ \sigma,\ \lambda$ - as follows :

$$p(r = 1|s; \mu, \sigma, \lambda) = \frac{1}{2} \cdot \lambda + (1 - \lambda) \cdot \Phi(s; \mu, \sigma), \tag{3}$$

where $r = 1$ stands for a "clockwise" response. The parameters are the bias or point of subjective equality (PSE) denoted with μ, the noise or inverse slope parameter denoted σ - which both are inputs to the Gaussian cumulative density function (Φ) - and the lapse rate, λ.

We first fitted a psychometric curve with 3 parameters for each one of the 4 conditions: No-Adapt-See, No-Adapt-Believe, Adapt-See and Adapt-Believe. However, model comparison selected a model with a a shared lapse λ across the 4 conditions based on AIC and BIC, so our main psychometric-curve cdf model had free parameters μ and σ for each of the 4 conditions, and a single free parameter λ shared across conditions (9 free parameters total; Fig. 3B).

Parameter estimation.

We performed maximum-likelihood estimation of the psychometric curve free parameters (4 μ, 4 σ, and 1 shared λ). The likelihood of a parameter combination is the probability of the data given that

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

parameter combination, with the log likelihood denoted as LL. We assumed that trials were independent and summed the log likelihoods across all trials. The LL for all the trials across all conditions is:

$$\begin{aligned}
&\text{LL}(\text{parameters}) = \log p(\text{data}|\text{parameters}) \\
&= \sum_{\text{No-Adapt-See trials } j} \log\ p(r_j|s_j; \mu_{\text{No-Adapt-See}},\ \sigma_{\text{No-Adapt-See}},\ \lambda) + \\
&\sum_{\text{No-Adapt-Believe trials } j} \log\ p(r_j|s_j; \mu_{\text{No-Adapt-Believe}},\ \sigma_{\text{No-Adapt-Believe}},\ \lambda) + \\
&\sum_{\text{Adapt-See trials } j} \log\ p(r_j|s_j; \mu_{\text{Adapt-See}},\ \sigma_{\text{Adapt-See}},\ \lambda) + \\
&\sum_{\text{Adapt-Believe trials } j} \log\ p(r_j|s_j; \mu_{\text{Adapt-Believe}},\ \sigma_{\text{Adapt-Believe}},\ \lambda)
\end{aligned}$$

where s_j and r_j are the stimulus and the participant's response on the j th trial, respectively.

Parameter estimation was performed via grid searches. For Experiment 1, we searched on a multidimensional grid with 101 values as follows: for each value of λ linearly spaced from 0 to 0.3, we performed 4 2-dimensional searches to find the corresponding μ and σ that maximized the LL per condition, with μ linearly spaced from -0.3 to 0.5 and σ logarithmically spaced from 0.001 and 0.5. We then found the global maximum LL across all the possible λ values and picked the corresponding set of 4 separate values of μ and σ. We used the same strategy for Experiment 2, but each parameter grid had 201 values instead. μ was linearly spaced from -0.2 to 0.2, σ was logarithmically spaced from 0.0001 and 0.2, and λ was linearly spaced from 0 to 0.3.

S2 Computational models

We now specify the standard perceptual decision-making model and use it to build the perceptual-insight model. As in [99, 100] and elsewhere, we first specify the generative or encoding models assumptions, then the decision rules, and then the generation of model predictions.

S2.1 Standard perceptual decision model

Generative model.

We present the generative model of the standard perceptual decision model in Fig. 5A, left. We denote by C the category of the spiral motion, which can take the values of $C = 1$ for clockwise or $C = -1$ for counterclockwise . The two categories occur about half the time each such that the correct prior would be: $p(C = 1) = p(C = -1) = \frac{1}{2}$. The stimulus s is the speed of motion of the test stimulus and takes a value in the interval [-0.3, 0.3], uniformly distributed: $s \sim \mathcal{U}(s; -0.3, 0.3)$ (similar to a case in [101]). The generative model further assumes that the observer's brain encodes a noisy measurement x of the stimulus s, which here we assume to be corrupted by Gaussian noise with standard deviation σ.

$$p(x|s) = \mathcal{N}(x; s, \sigma) \tag{4}$$

Generative model with distortion.

In the generative model with distortion (Fig. 5A, center), we assume the internal representation x is also offset by the factor A, with some value which we also call A. This value would plausibly

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

be very close to 0 in the No-Adapt conditions. In the Adapt conditions, we assume A is drawn from $N(A; \mu_A, \sigma_A)$. Thus, x depends on s, A and σ as follows:

$$p(x|s, A) = \mathcal{N}(x; s - A, \sigma) \tag{5}$$

Note that previous models of adaptation described two main features: the repulsion bias and increased sensitivity around the adaptor [61]. Here, we focused on the repulsion bias, or the offset in x induced by A. Increased sensitivity around the adaptor was not relevant to our data as the adaptor was much faster than the test stimuli and thus we could not measure the sensitivity around the adaptor.

Decision rule.

We assume that the observers compute the optimal decision rule to eventually generate their response for category $\hat{C}$ and confidence q (Fig. 1C). The inference process assumes that the Bayesian observer inverts the generative model to eventually produce their responses; it is depicted in Fig. 5A, right. As mentioned in the introduction, before producing their responses, we assume that the observer computes the posterior $p(C|x)$:

Using the decision variable, we can write the posterior:

$$p(C = 1|x) = \frac{p(C = 1|x)}{p(C = 1|x) + p(C = -1|x)} = \frac{1}{1 + \left(\frac{p(C=-1|x)}{p(C=1|x)}\right)} = \frac{1}{1 + e^{-d}} \tag{6}$$

Correspondingly, $p(C = -1|x) = \frac{1}{1+e^{d}}$.

We assume the observer picks the option $\hat{C}$ with the highest posterior (MAP):

$$\hat{C} = \underset{C}{\operatorname{argmax}}\, p(C|x) \tag{7}$$

This MAP decision rule is equivalent to reporting:

$$\hat{C} = \begin{cases} 1, & \text{if } d > 0 \\ -1, & \text{otherwise} \end{cases}$$

Instead of comparing d to 0, the observer may use a different criterion or threshold k_{choice} (Fig. 1C, center). While we do not delve into this here, the observer might perform the computation of the decision variable d in a way affected by decision noise, as assumed by [17] and others (originally [102]).

The observer also reports the confidence associated with their choice. We assume that the observer computes and reports confidence based on the Bayesian confidence hypothesis [16,17], where confidence q depends on the perceptual-decision variable and represents the posterior probability of the observer's choice, assumption in line with substantial empirical evidence and theoretical work [15,103–105]. As the reported confidence q was binary (low or high, coded respectively as 0 or 1), we assume that observers's confidence response are determined via a confidence threshold $k_{\text{confidence}}$, as:

$$\text{confidence} \quad = \quad p(C = \hat{C}|x) = \frac{1}{1 + e^{|d|}} \tag{8}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$q = \begin{cases} 1, & \text{if confidence} > k_{\text{confidence}} \\ 0, & \text{otherwise} \end{cases}$$

Alternative models of confidence have been proposed in the literature [106–109]. These include models with similar architecture to our basic model but which account for additional sources of noise and uncertainty (e.g., variability in the confidence criterion $k_{\text{confidence}}$ [110–113] or uncertainty about the confidence variable [114]). We opted for the basic, more established, architecture described above because it was more parsimonious yet appropriate for our design and since it provided a good description of our data.

Model predictions

While the decision rule prescribes how to jointly compute observer's responses $\hat{C}$ and subsequently confidence responses q when their internal measurement x is known, we emphasize that on every trial, the observers' internal measurement x is not known, but the test stimulus s is known. Thus, to find the predicted responses we need to compute the integral $p(\hat{C}, q|s)$. To do this, we have to marginalize over all the possible measurements x:

$$p(\hat{C}, q|s) = \int p(\hat{C}, q|x)\, p(x|s) dx \tag{9}$$

We computed the probability of responses $p(\hat{C}, q|s)$ by estimating the above integral by sampling: we simulated several samples (here $N_{\text{samples}} = 500$) of x from s and averaged the predictions over the corresponding outcomes $p(\hat{C}, q|x)$.

S2.2 Perceptual-insight model

Generative model with distortion.

The generative model is the exact same model as above in **Generative model with distortion**.

Decision rule. In the perceptual-insight model, the observer takes into account the correct generative model with the distortion factor A and then the computation of their posterior $p(C|x)$ is as follows:

$$p(C|x) \propto p(x|C)\, p(C) = p(C) \int p(x|C, A)\, p(A)\, dA = p(C) \iint p(x|s, A)\, p(s|C)\, p(A)\, ds\, dA \tag{10}$$

Below, we unpack the decision variable (which consists of the log-posterior ratio as above) using the distributional assumptions spelled out in the generative model with distortion:

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$
\begin{aligned}
d &= \log \frac{p(C=1|x)}{p(C=-1|x)} \\
&= \log \frac{p(C=1) \iint p(x|s,A)\, p(A)\, p(s|C=1)\, ds\, dA}{p(C=-1) \iint p(x|s,A)\, p(A)\, p(s|C=-1)\, ds\, dA} \\
&= \log \frac{p(C=1)}{p(C=-1)} \frac{\iint \mathcal{N}(x; s-A, \sigma)\, \mathcal{N}(A; \mu_A, \sigma_A)\, \mathcal{U}(s; s_{11}, s_{12})\, ds\, dA}{\iint \mathcal{N}(x; s-A, \sigma)\, \mathcal{N}(A; \mu_A, \sigma_A)\, \mathcal{U}(s; s_{21}, s_{22})\, ds\, dA} \\
&= \log \frac{p(C=1)}{p(C=-1)} \frac{\frac{1}{s_{12}-s_{11}} \iint \mathcal{N}(x; s-A, \sigma)\, \mathcal{N}(A; \mu_A, \sigma_A)\ ds\, dA}{\frac{1}{s_{22}-s_{21}} \iint \mathcal{N}(x; s-A, \sigma)\, \mathcal{N}(A; \mu_A, \sigma_A)\ ds\, dA} \qquad (11) \\
&= \log \frac{p(C=1)}{p(C=-1)} \frac{\frac{1}{s_{12}-s_{11}} \int_{-\infty}^{\infty} \int_{s=s_{11}}^{s=s_{12}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-(s-A))^2}{2\sigma^2}}\ \frac{1}{\sigma_A\sqrt{2\pi}} e^{-\frac{(A-\mu_A)^2}{2{\sigma_A}^2}}\ ds\, dA}{\frac{1}{s_{22}-s_{21}} \int_{-\infty}^{\infty} \int_{s=s_{21}}^{s=s_{22}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-(s-A))^2}{2\sigma^2}}\ \frac{1}{\sigma_A\sqrt{2\pi}} e^{-\frac{(A-\mu_A)^2}{2{\sigma_A}^2}}\ ds\, dA} \\
&= \log \frac{p(C=1)}{p(C=-1)} \frac{\frac{1}{s_{12}-s_{11}} \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{\sigma_A\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{s=s_{11}}^{s=s_{12}} e^{-\frac{(x-(s-A))^2}{2\sigma^2}}\ e^{-\frac{(A-\mu_A)^2}{2{\sigma_A}^2}}\ ds\, dA}{\frac{1}{s_{22}-s_{21}} \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{\sigma_A\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{s=s_{21}}^{s=s_{22}} e^{-\frac{(x-(s-A))^2}{2\sigma^2}}\ e^{-\frac{(A-\mu_A)^2}{2{\sigma_A}^2}}\ ds\, dA} \\
&= \log \frac{p(C=1)}{p(C=-1)} \frac{\frac{1}{s_{12}-s_{11}} \int_{-\infty}^{\infty} \int_{s=s_{11}}^{s=s_{12}} e^{-\frac{(s-(A+x))^2}{2\sigma^2}}\ e^{-\frac{(A-\mu_A)^2}{2{\sigma_A}^2}}\ ds\, dA}{\frac{1}{s_{22}-s_{21}} \int_{-\infty}^{\infty} \int_{s=s_{21}}^{s=s_{22}} e^{-\frac{(s-(A+x))^2}{2\sigma^2}}\ e^{-\frac{(A-\mu_A)^2}{2{\sigma_A}^2}}\ ds\, dA} \qquad (12)
\end{aligned}
$$

This integral can be solved numerically. However, for the models we test (Table S1), to make the calculations more tractable, we make the assumption that the observer has negligible uncertainty about their offset A, such that they have essentially no uncertainty over it (σ_A asymptotically close to 0). Nonetheless, the model had flexibility in this respect as it allowed σ_{encoding} parameters to vary in each condition, which could – and seemingly did – absorb additional uncertainty likely associated with A in the Adapt-Believe condition.

Normalized Gaussian functions with the width tending to 0, $\lim_{w\to 0}$, will yield Dirac delta functions δ:

$$\delta_w(x) = \lim_{w\to 0} \frac{1}{|w|\sqrt{\pi}} e^{-(\frac{x}{w})^2}$$

$$\delta(x) = \begin{cases} \infty, & \text{if } x = 0 \\ 0, & \text{otherwise} \end{cases}$$

The Delta function has the following property for any continuous function $f(x)$:

$$\int_{-\infty}^{\infty} f(x)\, \delta(x - x_0)\, dx = f(x_0)$$

Assuming that in $p(A)$ σ_A is asymptotically close to 0, we will get:

$$N(A; \mu_A, \sigma_A) \propto \frac{1}{|\sigma_A|\sqrt{\pi}} e^{-\left(\frac{A-\mu_A}{\sigma_A^2}\right)} = \delta_{\sigma_A}(A - \mu_A) \qquad (13)$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Substituting the relevant terms into Eq. 11 and using the property of the δ function we get:

$$\iint \mathcal{N}(x; s-A, \sigma)\, \mathcal{N}(A; \mu_A, \sigma_A)\; ds\, dA = \iint \mathcal{N}(x; s-A, \sigma)\, \delta_{\sigma_A}(A-\mu_A) ds\, dA = \int \mathcal{N}(x; s-\mu_A, \sigma) ds$$

Therefore:

$$\begin{aligned} d &= \log \frac{p(C=1)}{p(C=-1)} \frac{\int_{s_{11}}^{s_{12}} \mathcal{N}(x; s-\mu_A, \sigma)\, \mathcal{U}(s; s_{11}, s_{12})\, ds}{\int_{s_{11}}^{s_{12}} \mathcal{N}(x; s-\mu_A, \sigma)\, \mathcal{U}(s; s_{21}, s_{22})\, ds} \\ &= \log \frac{p(C=1)}{p(C=-1)} \frac{\frac{1}{s_{12}-s_{11}} \int_{s_{11}}^{s_{12}} \mathcal{N}(x; s-\mu_A, \sigma)\, ds}{\frac{1}{s_{22}-s_{21}} \int_{s_{21}}^{s_{22}} \mathcal{N}(x; s-\mu_A, \sigma)\, ds} \end{aligned} \tag{14}$$

The set of test stimuli in Experiment 2: $s_{11} = 0, s_{12} = 0.3$ and $s_{21} = -0.3, s_{22} = 0$. We make the change of variable $t = s - \mu_A$, with $dt = ds$, we substitute the limits of integration accordingly and write:

$$\begin{aligned} d &= \log \frac{p(C=1)}{p(C=-1)} \frac{\int_{s_{11}-\mu_A}^{s_{12}-\mu_A} \mathcal{N}(x; t, \sigma)\, dt}{\int_{s_{21}-\mu_A}^{s_{22}-\mu_A} \mathcal{N}(x; t, \sigma)\, dt} \\ &= \log \frac{p(C=1)}{p(C=-1)} \frac{\int_{s_{11}-\mu_A}^{s_{12}-\mu_A} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-t)^2}{2\sigma^2}}\, dt}{\int_{s_{21}-\mu_A}^{s_{22}-\mu_A} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-t)^2}{2\sigma^2}}\, dt} \\ &= \log \frac{p(C=1)}{p(C=-1)} \frac{\int_{s_{11}-\mu_A}^{s_{12}-\mu_A} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-x)^2}{2\sigma^2}}\, dt}{\int_{s_{21}-\mu_A}^{s_{22}-\mu_A} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-x)^2}{2\sigma^2}}\, dt} \\ &= \log \frac{p(C=1)}{p(C=-1)} \frac{\int_{s_{11}-\mu_A}^{s_{12}-\mu_A} \mathcal{N}(t; x, \sigma)\, dt}{\int_{s_{21}-\mu_A}^{s_{22}-\mu_A} \mathcal{N}(t; x, \sigma)\, dt} \end{aligned}$$

We use the property that : $Pr[La \leq x \leq Lb] = \int_{La}^{Lb} N(x; \mu, \sigma) dx = \frac{1}{2} \cdot \left[\text{erf}\left(\frac{Lb-\mu}{\sqrt{2}\,\sigma}\right) - \text{erf}\left(\frac{La-\mu}{\sqrt{2}\,\sigma}\right)\right]$

$$d = \log \frac{p(C=1)}{p(C=-1)} \frac{\frac{1}{2}\left[\text{erf}\left(\frac{s_{12}-\mu_A-x}{\sqrt{2}\,\sigma}\right) - \text{erf}\left(\frac{s_{11}-\mu_A-x}{\sqrt{2}\,\sigma}\right)\right]}{\frac{1}{2}\left[\text{erf}\left(\frac{s_{22}-\mu_A-x}{\sqrt{2}\,\sigma}\right) - \text{erf}\left(\frac{s_{21}-\mu_A-x}{\sqrt{2}\,\sigma}\right)\right]} \tag{15}$$

Because $\Phi(x; s, \sigma) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-s}{\sqrt{2}\,\sigma}\right)\right]$ and thus $\text{erf}\left(\frac{x-s}{\sqrt{2}\,\sigma}\right) = 2\Phi(x; s, \sigma) - 1$ we write:

$$\begin{aligned} d &= \log \frac{p(C=1)}{p(C=-1)} \frac{[2 * \Phi(s_{12} - \mu_A - x; 0, \sigma) - 1 - 2 * \Phi(s_{11} - \mu_A - x; 0, \sigma) + 1]}{[2 * \Phi(s_{22} - \mu_A - x; 0, \sigma) - 1 - 2 * \Phi(s_{21} - \mu_A - x; 0, \sigma) + 1]} \\ &= \log \frac{p(C=1)}{p(C=-1)} \frac{[\Phi(s_{12} - \mu_A - x; 0, \sigma) - \Phi(s_{11} - \mu_A - x; 0, \sigma)]}{[\Phi(s_{22} - \mu_A - x; 0, \sigma) - \Phi(s_{21} - \mu_A - x; 0, \sigma)]} \end{aligned}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We plug in the values $s_{11} = 0, s_{12} = 0.3$ and $s_{21} = -0.3, s_{22} = 0$, take the log and write out the decision variable:

$$d = \log \frac{p(C=1)}{1-p(C=1)} + \log \left(\frac{\Phi(0.3 - \mu_A - x; 0, \sigma) - \Phi(-\mu_A - x; 0, \sigma)}{\Phi(-\mu_A - x; 0, \sigma) - \Phi(-0.3 - \mu_A - x; 0, \sigma)} \right) \tag{16}$$

An insightful observer will know that they have an offset $\mu_A \neq 0$ and incorporate this into their decision variable. To distinguish between μ from the generative model and this value used in inference, we refer to them respectively as μ_{encoding} and $\mu_{\text{likelihood}}$. In the case of the Adapt conditions μ_{encoding} is offset. In Adapt-See, observers may ignore this offset and thus use $\mu_{\text{likelihood}} \approx 0$. In Adapt-Believe, instead, observers are asked to incorporate their knowledge of the offset and will thus try to use a $\mu_{\text{likelihood}}$ approximating their estimate of μ_{encoding}. Inference can be insightful even if $\mu_{\text{likelihood}} \neq \mu_{\text{encoding}}$, but the closer these values are, the more accurate the compensation will be for perceptual distortions.

Once the decision variable is computed as shown here, its use in the decision rule to yield choice $\hat{C}$ and confidence q is assumed to be as described above by the MAP decision rule in Equation 8.

Model predictions are again as in the previous section.

S2.3 Bayesian model variants

We tested the model variants in the table below. All models had 6 parameters for the early sensory-encoding stage (2 fixed for μ_{encoding}, 4 free σ parameters) and to allow for MAE compensation either had free parameters for intermediate-stage processes ($\mu_{\text{likelihood}}$ and/or category prior), or late-stage processes (k_{choice}), or both. All models also had free parameters for the confidence threshold $k_{\text{confidence}}$. The perceptual-insight model used for fitting only had free parameters for $\mu_{\text{likelihood}}$ at the intermediate inference stage. The critical competing model was able to compensate for the illusion based solely on changes in k_{choice} at a late choice stage. Other hybrid models were evaluated for completeness. Model comparison based on goodness of fit is presented in Fig. 5A and details on model fitting are provided below in Section S2.4.

Stage	Model				
	Perceptual insight "$\mu_{\textbf{likelihood}}$"	"**prior**"	"$\mu_{\textbf{likelihood}}$ **+ prior**"	Late compensation "$k_{\textbf{choice}}$"	"$\mu_{\textbf{likelihood}} + k_{\textbf{choice}}$"
Early	$\mu_{\text{encoding}} \times 2$ (fixed) $\sigma_{\text{encoding}} \times 4$	$\mu_{\text{encoding}} \times 2$ (fixed) $\sigma_{\text{encoding}} \times 4$	$\mu_{\text{encoding}} \times 2$ (fixed) $\sigma_{\text{encoding}} \times 4$	$\mu_{\text{encoding}} \times 2$ (fixed) $\sigma_{\text{encoding}} \times 4$	$\mu_{\text{encoding}} \times 2$ (fixed) $\sigma_{\text{encoding}} \times 4$
Intermediate	$\mu_{\text{likelihood}} \times 4$	prior $\times 4$	$\mu_{\text{likelihood}} \times 4$ prior $\times 4$		$\mu_{\text{likelihood}} \times 4$
Late	$k_{\text{confidence}} \times 4$	$k_{\text{confidence}} \times 4$	$k_{\text{confidence}} \times 4$	$k_{\text{confidence}} \times 4$ $k_{\text{choice}} \times 4$	$k_{\text{confidence}} \times 4$ $k_{\text{choice}} \times 4$

Table S1: **Bayesian model variants.** The parameters of each of the model variants fitted to the data. The parameters that uniquely vary for each model variant are highlighted in red. $\times$ 4 means one parameter for each condition: No-Adapt-See, No-Adapt-Believe, Adapt-See and Adapt-Believe. $\times$ 2 (fixed) means that the μ_{encoding} parameters are fixed to the values fitted from the psychometric curves for No-Adapt-See and respectively Adapt-See.

S2.4 Model fitting

We fitted the Bayesian model variants jointly to the choice and confidence data separately for each condition. Across conditions, we used the model fitting strategy described below. We fixed the two μ_{encoding}

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

parameters to the μ values fitted from the psychometric curves for the No-Adapt-See and Adapt-See conditions. We performed maximum-likelihood estimation (MLE) of the remaining parameters in the Bayesian models (with parameters as in Table S1). For a particular model, the likelihood of a set of parameters θ is the probability of the data given those parameters, $p(\text{data}|\theta)$. We denote the log likelihood with LL. We assumed that trials are independent of each other and thus we could sum the log likelihoods across all trials:

$$\text{LL}(\theta) \quad = \quad \log p(\text{data}|\theta) = \log \left(\prod_{j=1}^{N_{\text{trials}}} p(\text{responses}_j | s_j, \theta) \right) = \sum_{j=1}^{N_{\text{trials}}} \log p(\text{responses}_j | s_j, \theta)$$

We denote the subject's responses on the jth trial with responses_j above, which could be $\hat{C}_j = 1$ for clockwise or respectively $\hat{C}_j = 0$ for counterclockwise, and $q = 1$ for high confidence or $q = 0$ for low confidence.

We replaced extreme values (0 or 1) of $p(\hat{C}, q|s)$ with $\frac{1}{N_{\text{samples}}}$ for 0 and with $1 - \frac{1}{N_{\text{samples}}}$ for 1.

As in the "Model predictions" section, we approximated the $p(\text{responses}_j|s_j, \theta)$ through sampling. Even with 500 samples, the log likelihood can be considered noisy. To find the parameters θ that maximize $\text{LL}(\theta)$ we used an optimization method called Bayesian adaptive direct search (BADS) [115] that is especially suited for noisy functions. For each dataset and model, we ran BADS with 8 starting points and chose the best fitting parameters among those. Within BADS, we set the estimated noise size to 1 (`options.NoiseSize=1`).

The parameter ranges were $[\log(0.001), \log(0.15)]$ for σ, $[0.0001, 0.0099]$ for the prior right, $[0.50000, 0.99999]$ for $k_{\text{confidence}}$, $[-10, 10]$ for k_{choice} and $[-3, 6.5]$ for $\mu_{\text{likelihood}}$ For these parameters, we set the hard bounds equal to these plausible bounds, except: $[0.4, 0.6]$ for prior right and $[-5, 5]$ for k_{choice}. None of the participants had the parameters fitted to their upper or lower bounds, suggesting that these ranges were meaningful.

S2.4.1 Model comparison

We performed model comparison based on the Akaike Information Criterion (AIC) [116] and the Bayesian Information criterion (BIC) [117]. These metrics are defined as $\text{AIC} = -2\text{LL}^* + 2n_{\text{parameters}}$ and $\text{BIC} = -2\text{LL}^* + n_{\text{parameters}} \log n_{\text{trials}}$, respectively, where LL* is the maximum log likelihood and $n_{\text{parameters}}$ and n_{trials} the number of free parameters and respectively the number of trials.

S3 Supplementary Results

S3.1 Task and illusion comprehension checks

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

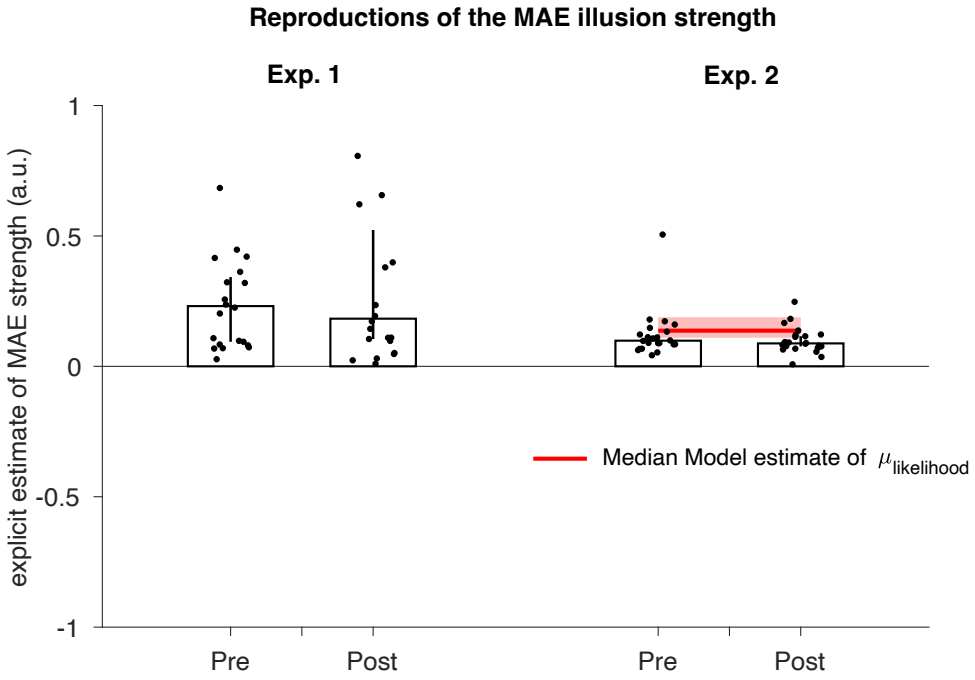

Figure S1: **Participants' explicit measures of MAE strength via trackpad and mouse reproductions provide evidence for MAE comprehension.** All points are above 0, showing that under an adaptor rotating to the left, participants estimated the illusion to be positive/ rotating to the right. For each participant, Pre shows the median of the first 10 estimates pre adaptation blocks and Post shows the median of the 10 estimates post the adaptation blocks. Bars depict medians across participants and bootstrapped 95% confidence intervals across 5000 samples. The red line and bar in Experiment 2 represent the $\mu_{\text{likelihood}}$ fitted values, again, median across participants and respectively bootstrapped 95% confidence intervals. Note how the participants' reproductions of MAE strength considerably overlap at the group level with the fitted $\mu_{\text{likelihood}}$ parameters for the Adapt-Believe condition from the perceptual-insight model. This suggests a similar magnitude at the group level of participants' explicit and implicit knowledge of the strength of their illusion. These values were not correlated with each other across participants; however, note that these illusion-reproduction estimates were based on longer adaptor exposures than the 3-s adaptor stimuli participants saw on each trial of the main experimental task, which precludes a direct comparison.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S3.2 Bias estimates from psychometric, confidence and reaction times are correlated

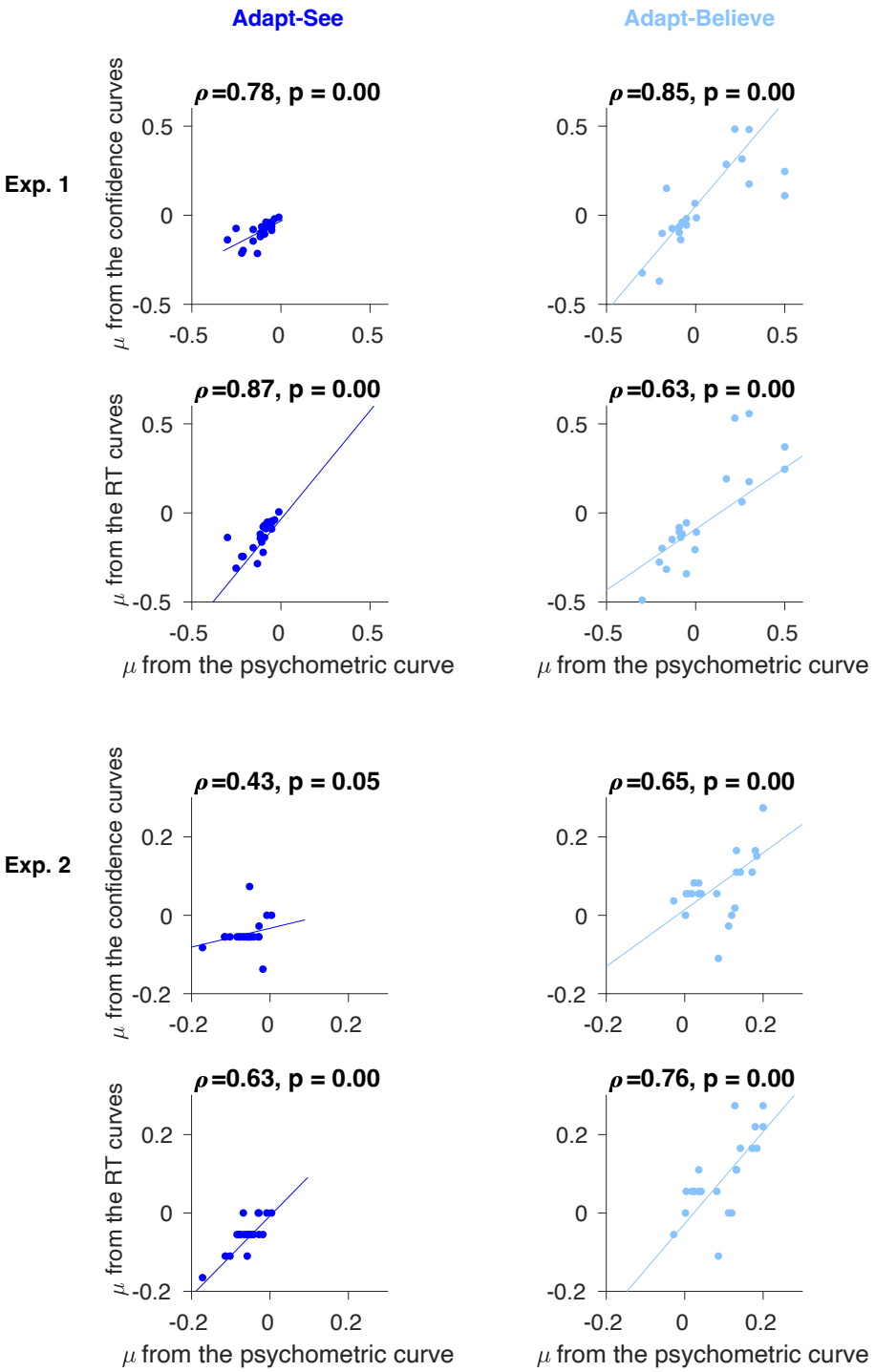

Figure S2: **Bias estimates from psychometric, confidence and reaction times are correlated.** We present in more details the locations of the peaks from Fig. 4. We see that these Spearman correlations are all significant across both Experiments 1 (Top) and 2 (Bottom), within both Adapt-See (Left) and Adapt-Believe (Right).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S3.3 More data from experiment 2: Psychometric, confidence and reaction time curves for 3 individual participants

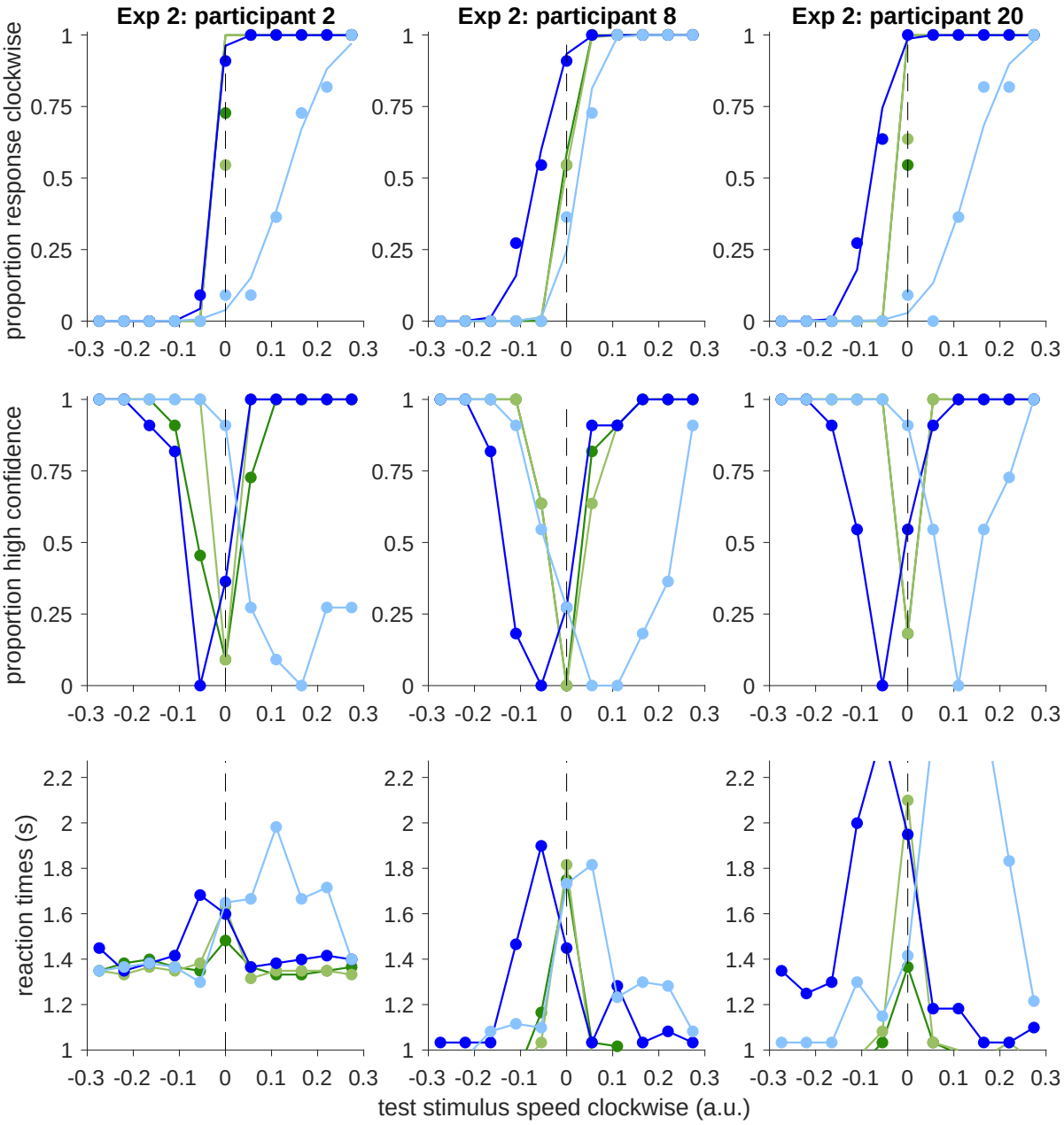

Figure S3: **Psychometric curves for MAE and MAE compensation shift in tandem with confidence and RT curves for 3 individual participants in Experiment 2.**

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S3.4 Bayesian models: simulation of how changes in parameters influence the decision variable, psychometric and confidence curves

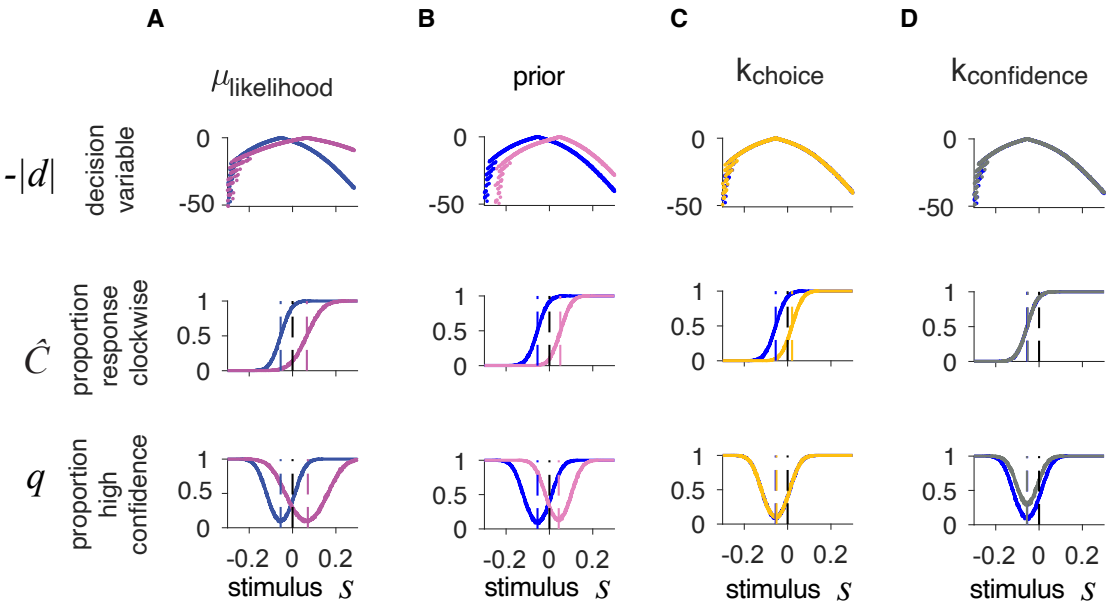

Figure S4: **Bayesian models: simulation of how changes in the parameters influence the decision variable, psychometric and confidence curves.** The parameters used to generate the blue curve are set approximately at the medians of the parameters inferred in the " winning model" in the Adapt-See conditions and are: $\mu_{\text{likelihood}} = 0$, $\mu_{\text{encoding}} = -0.055$, $\sigma = 0.04$, prior = 0.5, $k_{\text{choice}} = 0$, $k_{\text{confidence}} = 0.95$. Across the subplots, the curves with other colors show how changes in each parameter influence the simulated data as follows: **A)** $\mu_{\text{likelihood}} = -0.12$ **B)** prior = 0.1, **C)** $k_{\text{choice}} = 3.5$ and **D)** $k_{\text{confidence}} = 0.85$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S3.5 Experiment 2: Late-compensation model fits and parameters

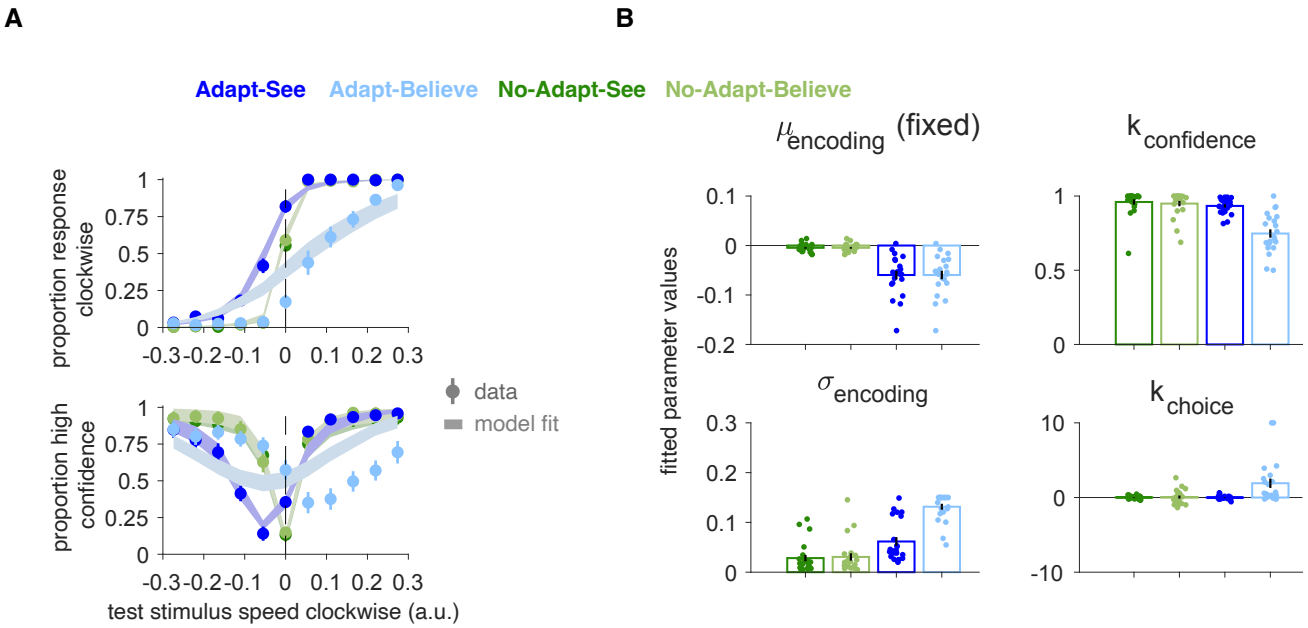

Figure S5: **Late-compensation or k_{choice} model fitted to the data from experiment 2**. **A)** k_{choice} model fits. **B)** Model parameters.

S3.6 Parameter recovery

To confirm that our parameter estimates are meaningful, we did a parameter recovery analysis for our winning $\mu_{\text{likelihood}}$ model. As above, we fixed $\mu_{\text{encoding}} = -0.055$ based on the median across participants for the psychometric curve fits from the Adapt-See conditions and as we fixed in the models for the Adapt-See and Adapt-Believe conditions. We varied $\mu_{\text{likelihood}}$, σ_{encoding} and $k_{\text{confidence}}$ across 6 relevant levels each (based on Fig. 5C) and for each parameter combination we simulated 3 datasets of 121 trials each (to match the number of trials we had in each experimental condition). Parameter recovery was successful, especially in the case of the parameter of interest $\mu_{\text{likelihood}}$.

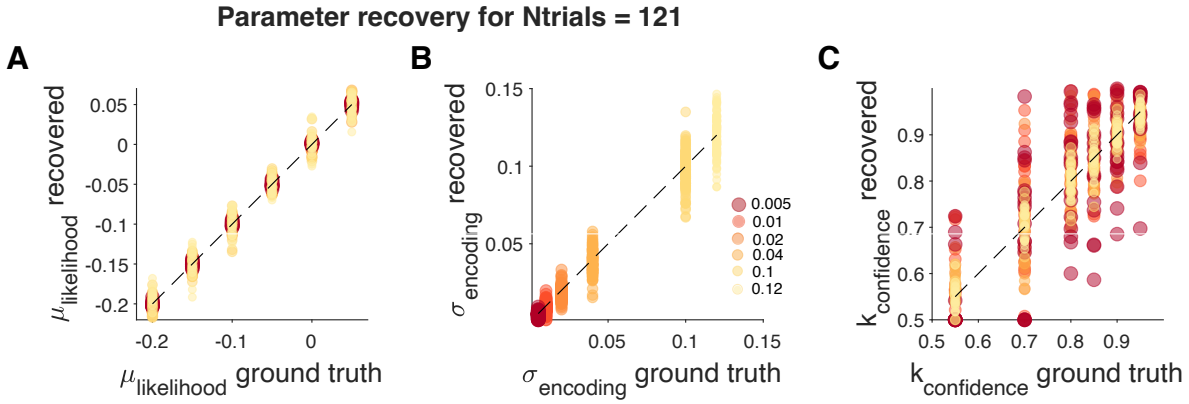

Figure S6: **Parameter recovery in the $\mu_{\text{likelihood}}$ model**. **A)** $\mu_{\text{likelihood}}$. **B)** σ_{encoding}. **C)** $k_{\text{confidence}}$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S3.7 Control experiment results

As expected, the observers performing the control experiment exhibited isolated shifts in the psychometric curves for Adapt-Bias relative to Adapt-See, in the absence of corresponding shifts in the confidence or RT curves (Fig S7A). We were able to fit our 5 models from S2.3 to 7/8 observers and found that the best fitting model was $\mu_{\text{likelihood}} + k_{\text{choice}}$ (AIC Median = 316, bootstrapped 95% CI: [218, 506], BIC Median = 372, bootstrapped 95% CI:[274, 561]).

On these data, the perceptual-insight model ($\mu_{\text{likelihood}}$) (median AIC = 368, median BIC = 409) and the response bias model (k_{choice}) (median AIC = 448, median BIC = 490) provided comparable fits ($p = 0.93$, two-tailed Wilcoxon signed-rank test on both AIC and BIC) (no z statistic available as the sample size was smaller than 15). The perceptual insight model fits were worse than the fits from the $\mu_{\text{likelihood}} + k_{\text{choice}}$ model ($p = 0.015$, left-tailed Wilcoxon signed-rank tests for AIC and BIC).

Critically, while the winning model $\mu_{\text{likelihood}} + k_{\text{choice}}$ contained both a the $\mu_{\text{likelihood}}$ and the k_{choice} parameters, $\mu_{\text{likelihood}}$ did not differ between Adapt-See and Adapt-Bias ($p = 0.47$), but k_{choice} did differ between these conditions ($p = 0.015$, two-tailed Wilcoxon signed-rank test) This supports the notion that a response-bias produces an isolated shift in psychometric curves that is distinct from intermediate inferential processes, and that this isolated shift is best captured as a late response-stage process via changes in k_{choice}.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

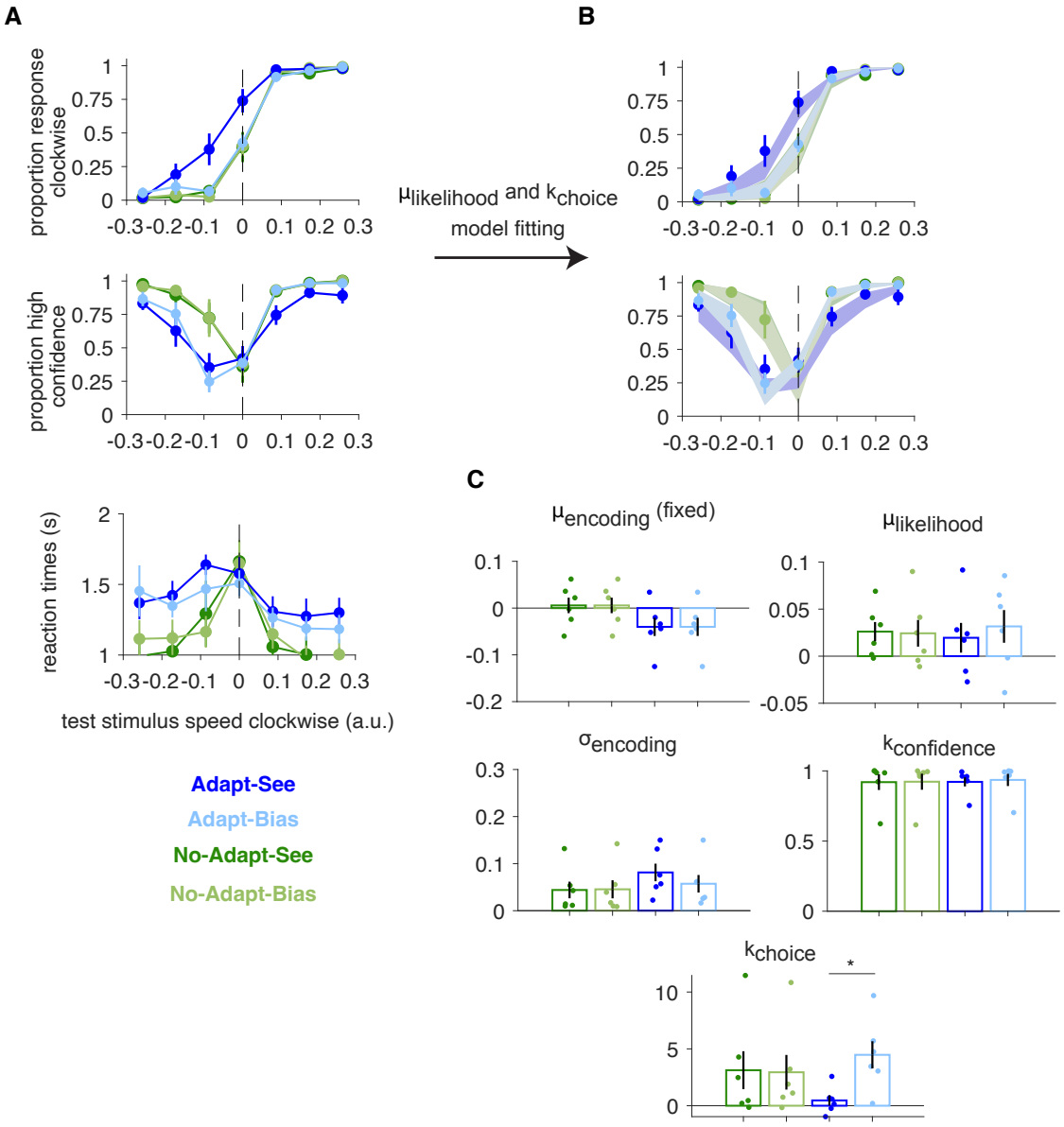

Figure S7: **Data from the pilot control experiment shows that a response-bias manipulation produces isolated shifts in the psychometric curves that are best captured by changes in k_{choice}. A)** Choice, confidence and reaction time data from N = 8 observers performing the control experiment. **B)** Fits from the $\mu_{\text{likelihood}} + k_{\text{choice}}$ winning model (shaded areas) capture well the choice and confidence curves (N = 7, mean $\pm$ SEM); **C)** Corresponding fitted parameter values (mean $\pm$ SEM). * indicates $p < 0.05$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements

We thank Kyo Iigaya, Mike Shadlen, Hakwan Lau, Daniel Wolpert, Amy Rapp, Ramon Nogueira and John Morrison for useful discussions and feedback on this project. We thank Ken Wengler, Brandon Ashinoff and Seth Baker for discussions and comments on this manuscript. We thank Alissa Fogelson, Anastasia Velikovskaya, Garrett Salzman and Jocelyn Kim for assistance with data collection. We also thank the participants for their time and work. This work was supported by the National Institute of Mental Health under awards R01MH117323 and R01MH114965 (GH).

References

[1] Maher, B. Delusional thinking and perceptual disorder. *The Journal of Individual Psychology* **30**, 98–113 (1974).

[2] Bentall, R. P. & Slade, P. D. Reality testing and auditory hallucinations: A signal detection analysis. *British Journal of Clinical Psychology* **24**, 159–169 (1985). URL https://doi.org/10.1111/j.2044-8260.1985.tb01331.x.

[3] Powers, A. R., Kelley, M. S. & Corlett, P. R. Varieties of voice-hearing: Psychics and the psychosis continuum. *Schizophrenia Bulletin* **43**, 84–98 (2016). URL https://doi.org/10.1093/schbul/sbw133.

[4] David, A. S. Insight and psychosis: the next 30 years. *The British Journal of Psychiatry* **217**, 521–523 (2019). URL https://doi.org/10.1192/bjp.2019.217.

[5] Acosta, F. J. Medication adherence in schizophrenia. *World Journal of Psychiatry* **2**, 74 (2012). URL https://doi.org/10.5498/wjp.v2.i5.74.

[6] Ljungdalh, P. Non-adherence to pharmacological treatment in schizophrenia and schizophrenia spectrum disorders – an updated systematic literature review. *The European Journal of Psychiatry* **31**, 172–186 (2017). URL https://doi.org/10.1016/j.ejpsy.2017.08.001.

[7] Vinogradov, S., Luks, T. L., Schulman, B. J. & Simpson, G. V. Deficit in a neural correlate of reality monitoring in schizophrenia patients. *Cerebral Cortex* **18**, 2532–2539 (2008). URL https://doi.org/10.1093/cercor/bhn028.

[8] Simons, J. S., Garrison, J. R. & Johnson, M. K. Brain mechanisms of reality monitoring. *Trends in Cognitive Sciences* **21**, 462–473 (2017). URL https://doi.org/10.1016/j.tics.2017.03.012.

[9] Winawer, J., Huk, A. C. & Boroditsky, L. A motion aftereffect from visual imagery of motion. *Cognition* **114**, 276–284 (2010). URL https://doi.org/10.1016/j.cognition.2009.09.010.

[10] Dijkstra, N., Mazor, M., Kok, P. & Fleming, S. Mistaking imagination for reality: Congruent mental imagery leads to more liberal perceptual detection. *Cognition* **212**, 104719 (2021). URL https://doi.org/10.1016/j.cognition.2021.104719.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[11] Dijkstra, N., Kok, P. & Fleming, S. M. Perceptual reality monitoring: Neural mechanisms dissociating imagination from reality. *Neuroscience & Biobehavioral Reviews* **135**, 104557 (2022). URL https://doi.org/10.1016/j.neubiorev.2022.104557.

[12] Dijkstra, N., Kok, P. & Fleming, S. M. Imagery adds stimulus-specific sensory evidence to perceptual detection. *Journal of Vision* **22**, 11 (2022). URL https://doi.org/10.1167/jov.22.2.11.

[13] Green, D. & Swets, J. *Signal detection theory and psychophysics* (John Wiley and Sons, 1966).

[14] Dayan, P. & Daw, N. D. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience* **8**, 429–453 (2008). URL https://doi.org/10.3758/cabn.8.4.429.

[15] Meyniel, F., Sigman, M. & Mainen, Z. F. Confidence as bayesian probability: From neural origins to behavior **88**, 78–92 (2015). URL https://doi.org/10.1016/j.neuron.2015.09.039.

[16] Adler, W. T. & Ma, W. J. Comparing bayesian and non-bayesian accounts of human confidence reports. *PLOS Computational Biology* **14**, e1006572 (2018). URL https://doi.org/10.1371/journal.pcbi.1006572.

[17] Li, H.-H. & Ma, W. J. Confidence reports in decision-making with multiple alternatives violate the bayesian confidence hypothesis. *Nature Communications* **11** (2020). URL https://doi.org/10.1038/s41467-020-15581-6.

[18] Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002). URL https://doi.org/10.1038/415429a.

[19] Jacobs, R. A. Optimal integration of texture and motion cues to depth. *Vision Research* **39**, 3621–3629 (1999). URL https://doi.org/10.1016/s0042-6989(99)00088-7.

[20] Körding, K. P. *et al.* Causal inference in multisensory perception. *PLoS ONE* **2**, e943 (2007). URL https://doi.org/10.1371/journal.pone.0000943.

[21] Petzschner, F. H., Maier, P. & Glasauer, S. Combining symbolic cues with sensory input and prior experience in an iterative bayesian framework. *Frontiers in Integrative Neuroscience* **6** (2012). URL https://doi.org/10.3389/fnint.2012.00058.

[22] Mendonca, A. G. *et al.* The impact of learning on perceptual decisions and its implication for speed-accuracy tradeoffs. *Nature Communications* **11** (2020). URL https://doi.org/10.1038/s41467-020-16196-7.

[23] Haddara, N. & Rahnev, D. The impact of feedback on perceptual decision-making and metacognition: Reduction in bias but no change in sensitivity. *Psychological Science* **33**, 259–275 (2022). URL https://doi.org/10.1177/09567976211032887.

[24] Gallagher, R. M., Suddendorf, T. & Arnold, D. H. Confidence as a diagnostic tool for perceptual aftereffects. *Scientific Reports* **9** (2019). URL https://doi.org/10.1038/s41598-019-43170-1.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[25] Moscoso, P. A. M., Cicchini, G. M., Arrighi, R. & Burr, D. C. Adaptation to hand-tapping affects sensory processing of numerosity directly: evidence from reaction times and confidence. *Proceedings of the Royal Society B: Biological Sciences* **287**, 20200801 (2020). URL https://doi.org/10.1098/rspb.2020.0801.

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

[27] Steiner, V., Blake, R. & Rose, D. Interocular transfer of expansion, rotation, and translation motion aftereffects. *Perception* **23**, 1197–1202 (1994). URL https://doi.org/10.1068/p231197.

[28] Bex, P. J., Metha, A. B. & Makous, W. Enhanced motion aftereffect for complex motions. *Vision Research* **39**, 2229–2238 (1999). URL https://doi.org/10.1016/s0042-6989(98)00329-0.

[29] Blake, R. & Hiris, E. Another means for measuring the motion aftereffect. *Vision Research* **33**, 1589–1592 (1993). URL https://doi.org/10.1016/0042-6989(93)90152-m.

[30] Clifford, C. W. G. & Rhodes, G. *Fitting the Mind to the World: Adaptation and After-Effects in High-Level Vision* (Oxford University Press, 2005). URL https://doi.org/10.1093/acprof:oso/9780198529699.001.0001.

[31] Waskom, M. L., Okazawa, G. & Kiani, R. Designing and interpreting psychophysical investigations of cognition. *Neuron* **104**, 100–112 (2019). URL https://doi.org/10.1016/j.neuron.2019.09.016.

[32] Morgan, M., Dillenburger, B., Raphael, S. & Solomon, J. A. Observers can voluntarily shift their psychometric functions without losing sensitivity. *Attention, Perception, & Psychophysics* **74**, 185–193 (2011). URL https://doi.org/10.3758/s13414-011-0222-7.

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

[34] Stuit, S. M. Motion processing, adaptation and aftereffects: A review (2009).

[35] van Wezel, R. & Britten, K. Multiple uses of visual motion. the case for stability in sensory cortex. *Neuroscience* **111**, 739–759 (2002). URL https://doi.org/10.1016/s0306-4522(02)00036-2.

[36] Carandini, M., Movshon, J. A. & Ferster, D. Pattern adaptation and cross-orientation interactions in the primary visual cortex. *Neuropharmacology* **37**, 501–511 (1998). URL https://doi.org/10.1016/s0028-3908(98)00069-0.

[37] Graziano, M., Andersen, R. & Snowden, R. Tuning of MST neurons to spiral motions. *The Journal of Neuroscience* **14**, 54–67 (1994). URL https://doi.org/10.1523/jneurosci.14-01-00054.1994.

[38] Kohn, A. Visual adaptation: Physiology, mechanisms, and functional benefits. *Journal of Neurophysiology* **97**, 3155–3164 (2007). URL https://doi.org/10.1152/jn.00086.2007.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[39] Joshi, S. & Gold, J. I. Pupil size as a window on neural substrates of cognition. *Trends in Cognitive Sciences* **24**, 466–480 (2020). URL https://doi.org/10.1016/j.tics.2020.03.005.

[40] Allen, M. *et al.* Unexpected arousal modulates the influence of sensory noise on confidence. *eLife* **5** (2016). URL https://doi.org/10.7554/elife.18103.

[41] Tortelli, C., Turi, M., Burr, D. C. & Binda, P. Pupillary responses obey emmert's law and covary with autistic traits. *Journal of Autism and Developmental Disorders* (2020). URL https://doi.org/10.1007/s10803-020-04718-7.

[42] Satterthwaite, T. D. *et al.* Dissociable but inter-related systems of cognitive control and reward during decision making: Evidence from pupillometry and event-related fMRI. *NeuroImage* **37**, 1017–1031 (2007). URL https://doi.org/10.1016/j.neuroimage.2007.04.066.

[43] Zénon, A. Eye pupil signals information gain. *Proceedings of the Royal Society B: Biological Sciences* **286**, 20191593 (2019). URL https://doi.org/10.1098/rspb.2019.1593.

[44] Lempert, K. M., Chen, Y. L. & Fleming, S. M. Relating pupil dilation and metacognitive confidence during auditory decision-making. *PLOS ONE* **10**, e0126588 (2015). URL https://doi.org/10.1371/journal.pone.0126588.

[45] Graves, J. E., Egré, P., Pressnitzer, D. & de Gardelle, V. An implicit representation of stimulus ambiguity in pupil size. *Proceedings of the National Academy of Sciences* **118**, e2107997118 (2021). URL https://doi.org/10.1073/pnas.2107997118.

[46] Pitcher, G. *A Theory of Perception* (Princeton University Press, 1971).

[47] Fletcher, P. C. & Frith, C. D. Perceiving is believing: a bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience* **10**, 48–58 (2008). URL https://doi.org/10.1038/nrn2536.

[48] Lau, H. Consciousness, metacognition, and perceptual reality monitoring. *psyArXiv* (2019). URL https://doi.org/10.31234/osf.io/ckbyf.

[49] Gershman, S. J. The generative adversarial brain. *Frontiers in Artificial Intelligence* **2** (2019). URL https://doi.org/10.3389/frai.2019.00018.

[50] Turner, M. & Coltheart, M. Confabulation and delusion: A common monitoring framework. *Cognitive Neuropsychiatry* **15**, 346–376 (2010). URL https://doi.org/10.1080/13546800903441902.

[51] Fleming, S. M. & Lau, H. C. How to measure metacognition. *Frontiers in Human Neuroscience* **8** (2014). URL https://doi.org/10.3389/fnhum.2014.00443.

[52] Kohn, A. & Movshon, J. Neuronal adaptation to visual motion in area MT of the macaque. *Neuron* **39**, 681–691 (2003). URL https://doi.org/10.1016/s0896-6273(03)00438-0.

[53] Huk, A. C., Ress, D. & Heeger, D. J. Neuronal basis of the motion aftereffect reconsidered. *Neuron* **32**, 161–172 (2001). URL https://doi.org/10.1016/s0896-6273(01)00452-4.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[54] Fetsch, C. R., Kiani, R., Newsome, W. T. & Shadlen, M. N. Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron* **83**, 797–804 (2014). URL https://doi.org/10.1016/j.neuron.2014.07.011.

[55] Fetsch, C. R. *et al.* Focal optogenetic suppression in macaque area MT biases direction discrimination and decision confidence, but only transiently. *eLife* **7** (2018). URL https://doi.org/10.7554/elife.36523.

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

[57] Hanks, T. D., Ditterich, J. & Shadlen, M. N. Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. *Nature Neuroscience* **9**, 682–689 (2006). URL https://doi.org/10.1038/nn1683.

[58] Hanks, T. D. *et al.* Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature* **520**, 220–223 (2015). URL https://doi.org/10.1038/nature14066.

[59] Jazayeri, M. & Movshon, J. A. A new perceptual illusion reveals mechanisms of sensory decoding. *Nature* **446**, 912–915 (2007). URL https://doi.org/10.1038/nature05739.

[60] Stocker, A. & Simoncelli, E. Visual motion aftereffects arise from a cascade of two isomorphic adaptation mechanisms. *Journal of Vision* **9**, 9–9 (2009). URL https://doi.org/10.1167/9.9.9.

[61] Stocker, A. & Simoncelli, E. Sensory adaptation within a bayesian framework for perception. *Advances in Neural Information Processing Systems* 1291–1298 (2006).

[62] Seriès, P., Stocker, A. A. & Simoncelli, E. P. Is the homunculus "aware" of sensory adaptation? *Neural Computation* **21**, 3271–3304 (2009). URL https://doi.org/10.1162/neco.2009.09-08-869.

[63] Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nature Neuroscience* **9**, 1432–1438 (2006). URL https://doi.org/10.1038/nn1790.

[64] Beck, J. M. *et al.* Probabilistic population codes for bayesian decision making. *Neuron* **60**, 1142–1152 (2008). URL https://doi.org/10.1016/j.neuron.2008.09.021.

[65] Gallagher, R. M., Suddendorf, T. & Arnold, D. H. The implied motion aftereffect changes decisions, but not confidence. *Attention, Perception, & Psychophysics* **83**, 3047–3055 (2021). URL https://doi.org/10.3758/s13414-021-02331-z.

[66] Powers, A. R., Mathys, C. & Corlett, P. R. Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science* **357**, 596–600 (2017). URL https://doi.org/10.1126/science.aan3458.

[67] Corlett, P. R. *et al.* Hallucinations and strong priors. *Trends in Cognitive Sciences* **23**, 114–127 (2019). URL https://doi.org/10.1016/j.tics.2018.12.001.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[68] Ashinoff, B. K., Singletary, N. M., Baker, S. C. & Horga, G. Rethinking delusions: A selective review of delusion research through a computational lens. *Schizophrenia Research* (2021). URL https://doi.org/10.1016/j.schres.2021.01.023.

[69] Pavlova, B. & Uher, R. Assessment of psychopathology. *JAMA Psychiatry* **77**, 557 (2020). URL https://doi.org/10.1001/jamapsychiatry.2020.0108.

[70] Brown, H. E., Feinstein, S. & Henderson, D. C. Working with patients with impaired insight. *JAMA Psychiatry* **77**, 766 (2020). URL https://doi.org/10.1001/jamapsychiatry.2020.0767.

[71] Thakkar, K. N., Antinori, A., Carter, O. L. & Brascamp, J. W. Altered short-term neural plasticity related to schizotypal traits: Evidence from visual adaptation. *Schizophrenia Research* **207**, 48–57 (2018). URL https://doi.org/10.1016/j.schres.2018.04.013.

[72] Thakkar, K. N., Silverstein, S. M. & Brascamp, J. W. A review of visual aftereffects in schizophrenia. *Neuroscience & Biobehavioral Reviews* **101**, 68–77 (2019). URL https://doi.org/10.1016/j.neubiorev.2019.03.021.

[73] Thakkar, K. N. *et al.* Stronger tilt aftereffects in persons with schizophrenia. *Journal of Abnormal Psychology* **130**, 186–197 (2021). URL https://doi.org/10.1037/abn0000653.

[74] Scarfe, P. Spiral texture demo for matlab. URL https://peterscarfe.com/spiraltexturedemo.html.

[75] Powell, G., Sumner, P. & Bompas, A. The effect of eye movements and blinks on afterimage appearance and duration. *Journal of Vision* **15**, 20 (2015). URL https://doi.org/10.1167/15.3.20.

[76] Brainard, D. H. The psychophysics toolbox. *Spatial Vision* **10**, 433–436 (1997).

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

[78] Kleiner, M. *et al.* What's new in psychtoolbox-3. *Perception* **36**, 1–16 (2007).

[79] Kaunitz, L., Fracasso, A. & Melcher, D. Unseen complex motion is modulated by attention and generates a visible aftereffect. *Journal of Vision* **11**, 10–10 (2011). URL https://doi.org/10.1167/11.13.10.

[80] Cavanagh, P. & Favreau, O. E. Motion aftereffect: A global mechanism for the perception of rotation. *Perception* **9**, 175–182 (1980). URL https://doi.org/10.1068/p090175.

[81] Freeman, J., Heeger, D. J. & Merriam, E. P. Coarse-scale biases for spirals and orientation in human visual cortex. *Journal of Neuroscience* **33**, 19695–19703 (2013). URL https://doi.org/10.1523/jneurosci.0889-13.2013.

[82] Georgiades, M. & Harris, J. Evidence for spatio-temporal selectivity in attentional modulation of the motion aftereffect. *Spatial Vision* **16**, 21–31 (2002).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[83] Ashida, H. & Osaka, N. Motion aftereffect with flickering test stimuli depends on adapting velocity. *Vision Research* **35**, 1825–1833 (1995). URL https://doi.org/10.1016/0042-6989(94)00270-v.

[84] Keck, M. J., Palella, T. D. & Pantle, A. Motion aftereffect as a function of the contrast of sinusoidal gratings. *Vision Research* **16**, 187–191 (1976). URL https://doi.org/10.1016/0042-6989(76)90097-3.

[85] Nishida, S., Ashida, H. & Sato, T. Contrast dependencies of two types of motion aftereffect. *Vision Research* **37**, 553–563 (1997). URL https://doi.org/10.1016/s0042-6989(96)00181-2.

[86] Aghdaee, S. M. Adaptation to spiral motion in crowding condition. *Perception* **34**, 155–162 (2005). URL https://doi.org/10.1068/p5298.

[87] Acerbi, L. Bayesian adaptive stimulus placement of psychometric function for matlab (2016).

[88] Kontsevich, L. L. & Tyler, C. W. Bayesian adaptive estimation of psychometric slope and threshold. *Vision research* **39**, 2729–2737 (1999).

[89] Prins, N. The adaptive psi method and the lapse rate. *Journal of Vision* **12**, 322–322 (2012).

[90] Prins, N. The psi-marginal adaptive method: How to give nuisance parameters the attention they deserve (no more, no less). *Journal of Vision* **13**, 3–3 (2013). URL https://doi.org/10.1167/13.7.3.

[91] Mihali, A., Young, A. G., Adler, L. A., Halassa, M. M. & Ma, W. J. A low-level perceptual correlate of behavioral and clinical deficits in ADHD. *Computational Psychiatry* **2**, 141 (2018). URL https://doi.org/10.1162/cpsy_a_00018.

[92] Witthoft, N., Sha, L., Winawer, J. & Kiani, R. Sensory and decision-making processes underlying perceptual adaptation. *Journal of Vision* **18**, 10 (2018). URL https://doi.org/10.1167/18.8.10.

[93] Zokaei, N., Board, A. G., Manohar, S. G. & Nobre, A. C. Modulation of the pupillary response by the content of visual working memory. *Proceedings of the National Academy of Sciences* **116**, 22802–22810 (2019). URL https://doi.org/10.1073/pnas.1909959116.

[94] Cornelissen, F. W., Peters, E. M. & Palmer, J. The eyelink toolbox: Eye tracking with MATLAB and the psychophysics toolbox. *Behavior Research Methods, Instruments, & Computers* **34**, 613–617 (2002).

[95] Kret, M. E. & Sjak-Shie, E. E. Preprocessing pupil size data: Guidelines and code. *Behavior Research Methods* **51**, 1336–1342 (2018). URL https://doi.org/10.3758/s13428-018-1075-y.

[96] Mathôt, S., Fabius, J., Heusden, E. V. & der Stigchel, S. V. Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods* **50**, 94–106 (2018). URL https://doi.org/10.3758/s13428-017-1007-2.

[97] Binda, P., Pereverzeva, M. & Murray, S. O. Pupil constrictions to photographs of the sun. *Journal of Vision* **13**, 8–8 (2013). URL https://doi.org/10.1167/13.6.8.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[98] Wichmann, F. A. & Hill, N. J. The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & Psychophysics* **63**, 1293–1313 (2001).

[99] Ma, W. J. Bayesian decision models: A primer. *Neuron* **104**, 164–175 (2019). URL https://doi.org/10.1016/j.neuron.2019.09.037.

[100] Mihali, A. & Ma, W. J. The psychophysics of visual search with heterogeneous distractors. *bioRxiv* (2020). URL https://doi.org/10.1101/2020.08.10.244707.

[101] Adler, W. T. & Ma, W. J. Limitations of proposed signatures of bayesian confidence. *Neural Computation* **30**, 3327–3354 (2018). URL https://doi.org/10.1162/neco_a_01141.

[102] Mueller, S. T. & Weidemann, C. T. Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review* **15**, 465–494 (2008). URL https://doi.org/10.3758/pbr.15.3.465.

[103] Kiani, R. & Shadlen, M. N. Representation of confidence associated with a decision by neurons in the parietal cortex **324**, 759–764 (2009). URL https://doi.org/10.1126/science.1169405.

[104] Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals **19**, 366–374 (2016). URL https://doi.org/10.1038/nn.4240.

[105] Ma, W. J. & Jazayeri, M. Neural coding of uncertainty and probability **37**, 205–220 (2014). URL https://doi.org/10.1146/annurev-neuro-071013-014017.

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

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

[108] Rahnev, D. Visual metacognition: Measures, models, and neural correlates. *American Psychologist* **76**, 1445–1453 (2021). URL https://doi.org/10.1037/amp0000937.

[109] Rahnev, D. *et al.* Consensus goals in the field of visual metacognition. *Perspectives on Psychological Science* 174569162210756 (2022). URL https://doi.org/10.1177/17456916221075615.

[110] Martino, B. D., Fleming, S. M., Garrett, N. & Dolan, R. J. Confidence in value-based choice. *Nature Neuroscience* **16**, 105–110 (2012). URL https://doi.org/10.1038/nn.3279.

[111] Jang, Y., Wallsten, T. S. & Huber, D. E. A stochastic detection and retrieval model for the study of metacognition. *Psychological Review* **119**, 186–200 (2012). URL https://doi.org/10.1037/a0025960.

[112] van den Berg, R., Yoo, A. H. & Ma, W. J. Fechner's law in metacognition: A quantitative model of visual working memory confidence. *Psychological Review* **124**, 197–214 (2017). URL https://doi.org/10.1037/rev0000060.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.13.468497; this version posted September 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[113] Shekhar, M. & Rahnev, D. The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review* **128**, 45–70 (2021). URL https://doi.org/10.1037/rev0000249.

[114] Boundy-Singer, Z. M., Ziemba, C. M. & Goris, R. L. T. Confidence as a noisy decision reliability estimate. *bioRxiv* (2021). URL https://doi.org/10.1101/2021.12.17.473249.

[115] Acerbi, L. & Ma, W. J. Practical bayesian optimization for model fitting with bayesian adaptive direct search. *Advances in Neural Information Processing Systems* 1834–1844 (2017). URL https://proceedings.neurips.cc/paper/2017/file/df0aab058ce179e4f7ab135ed4e641a9-Paper.pdf.

[116] Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723 (1974). URL https://doi.org/10.1109/tac.1974.1100705.

[117] Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics* **6** (1978). URL https://doi.org/10.1214/aos/1176344136.