

# Beliefs About Perception Shape Perceptual Inference: An Ideal Observer Model of Detection

Matan Mazor<sup>1, 2, 3, 4</sup>, Rani Moran<sup>5, 6</sup>, and Clare Press<sup>3, 4, 7</sup>

<sup>1</sup> All Souls College, University of Oxford

<sup>2</sup> Department of Experimental Psychology, University of Oxford

<sup>3</sup> School of Psychological Sciences, Birkbeck, University of London

<sup>4</sup> Wellcome Centre for Human Neuroimaging, University College London

<sup>5</sup> School of Biological and Behavioural Sciences, Queen Mary University of London

<sup>6</sup> Max Planck University College London Centre for Computational Psychiatry and Aging Research

<sup>7</sup> Department of Experimental Psychology, University College London

According to Bayesian, “inverse optics” accounts of vision, perceiving is inferring the most likely state of the world given noisy sensory data. This inference depends not only on prior beliefs about the world but also on an internal model specifying how world states translate to visual sensations. Alternative accounts explain perceptual decisions as a rule-based process, with no role for such beliefs about perception. Here, we contrast the two alternatives, focusing on decisions about perceptual absence as a critical test case. We present data from three preregistered experiments where participants performed a near-threshold detection task under different levels of partial stimulus occlusion, thereby visibly manipulating the measurement function going from external world states to internal perceptual states. We find that decisions about presence and absence are differentially sensitive to sensory evidence and occlusion. Furthermore, we observe reliably opposite individual-level effects of occlusion on decisions about absence. Our model accounts for these findings by postulating robust individual differences in the incorporation of beliefs about visibility into perceptual inferences, independent of population variability in visibility itself. We discuss implications for the varied and inferential nature of visual perception more broadly.

*Keywords:* absence, probabilistic reasoning, perception

After checking Taylor Swift’s Wikipedia page, we are confident that she has not announced her retirement from music. If she had, it would have been mentioned on her page. We also checked cellist Natalia Gutman’s page and did not see any mention of a similar announcement, but we are not so sure she has not made one, since her Wikipedia page only gets updated irregularly. The absence of evidence on Wikipedia is enough to make a solid inference in the case of Swift but not in the case of Gutman, because we know that information about Swift spreads more efficiently on the Internet. In other words, we believe that something is not true when we believe that “if it were true, we would have heard about it” (Goldberg, 2011).

More generally, for a decision to be rational, it should depend not only on observation and prior beliefs but also on our beliefs

regarding the likelihood of the observation in hand given competing hypotheses about the world. This role for beliefs about the evidence-generating model is especially pertinent when the observation in question is that no evidence is available, as the absence of evidence for a signal can reflect the true absence of a signal or a failure to obtain evidence for it (Altman & Bland, 1995; Locke, 1690). This leaves a critical role for the decision-maker’s internal model of how likely evidence is to become available when a signal is present (Oaksford & Hahn, 2004; Walton, 1992, 2010).

According to inferential, “inverse optics” accounts of vision, a similar principle is at play in perception too. Such accounts describe seeing as inferring the most likely state of the world to have given rise to the observed sense data (see Figure 1, Friston, 2010;

---

Hongjing Lu served as action editor.

Matan Mazor  <https://orcid.org/0000-0002-3601-0644>

This work was supported by a European Research Council consolidator grant (Grant no. 101001592) under the European Union’s Horizon 2020 research and innovation programme, awarded to Clare Press. Matan Mazor is supported by a postdoctoral research fellowship from All Souls College at the University of Oxford. The authors thank Daniel Yon and Mathias Sablé-Meyer for useful feedback on previous versions of this article.

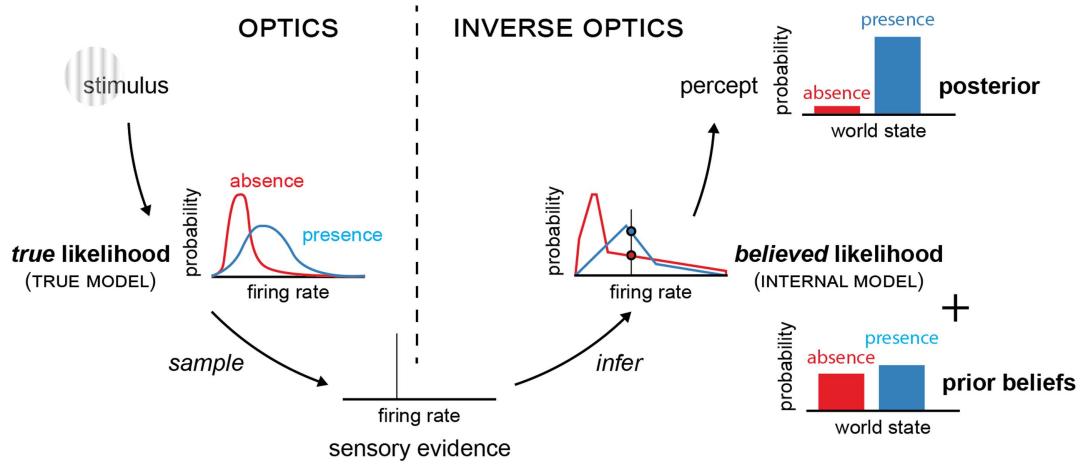
Open Access funding provided by University of Oxford: This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0; <https://creativecommons.org/licenses/by/4.0/>). This license

permits copying and redistributing the work in any medium or format, as well as adapting the material for any purpose, even commercially.

Matan Mazor played a lead role in conceptualization, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualization, writing—original draft, and writing—review and editing. Rani Moran played an equal role in formal analysis, software, and writing—review and editing. Clare Press played a lead role in funding acquisition, resources, and supervision and an equal role in conceptualization, investigation, and writing—review and editing.

Correspondence concerning this article should be addressed to Matan Mazor, All Souls College, University of Oxford, High Street, OX1 4AL Oxford, United Kingdom. Email: matan.mazor@psy.ox.ac.uk

**Figure 1**  
A Schematic of an Inverse Optics Account of Vision



*Note.* According to inverse optics accounts, vision is the process by which observers invert an internal model of optics to identify the most likely state of the world to have given rise to a piece of noisy evidence. Conditioned on a stimulus (top left), evidence (here, firing rate) is probabilistically sampled, following the true likelihood function. The agent infers the most likely stimulus by consulting a believed likelihood function, which is derived from a simplified model of their visual system, to extract a likelihood ratio (represented as the relative height of the blue and red dots). The likelihood ratio is integrated with prior beliefs about the world, resulting in a percept (posterior over world states; top right). See the online article for the color version of this figure.

Gershman et al., 2012; Smith, 2001; von Helmholtz, 1866). This renders percepts dependent on how the sensory system maps world states to sense data, and a subjective prior, indicating what the observer expects to be true about the world (Ma, 2019; Maloney & Mamassian, 2009; Mamassian et al., 2002; Press et al., 2020; Seth, 2014; Yon & Frith, 2021). Our focus here is on a third critical component of “inverse optics” accounts of vision: the observer’s internal model of the evidence-generating process by which its sensory system maps world states to sense data. This model is used to compare the likelihood of sensory samples under different world states (Herce Castañón et al., 2019; Ma, 2012). Critically, it need not be represented in declarative form, but can be implicitly encoded within the visual system itself. In the case of perceptual detection, the true model determines the objective probability that a stimulus would be registered in sensory channels, and the internal model determines whether the agent believes, explicitly or implicitly, that a stimulus would be registered in sensory channels, if present. We term this model-derived belief about the visibility of hypothetical stimuli “*expected visibility*,” and contrast it with *visibility*, which is dependent on the sensory system itself.

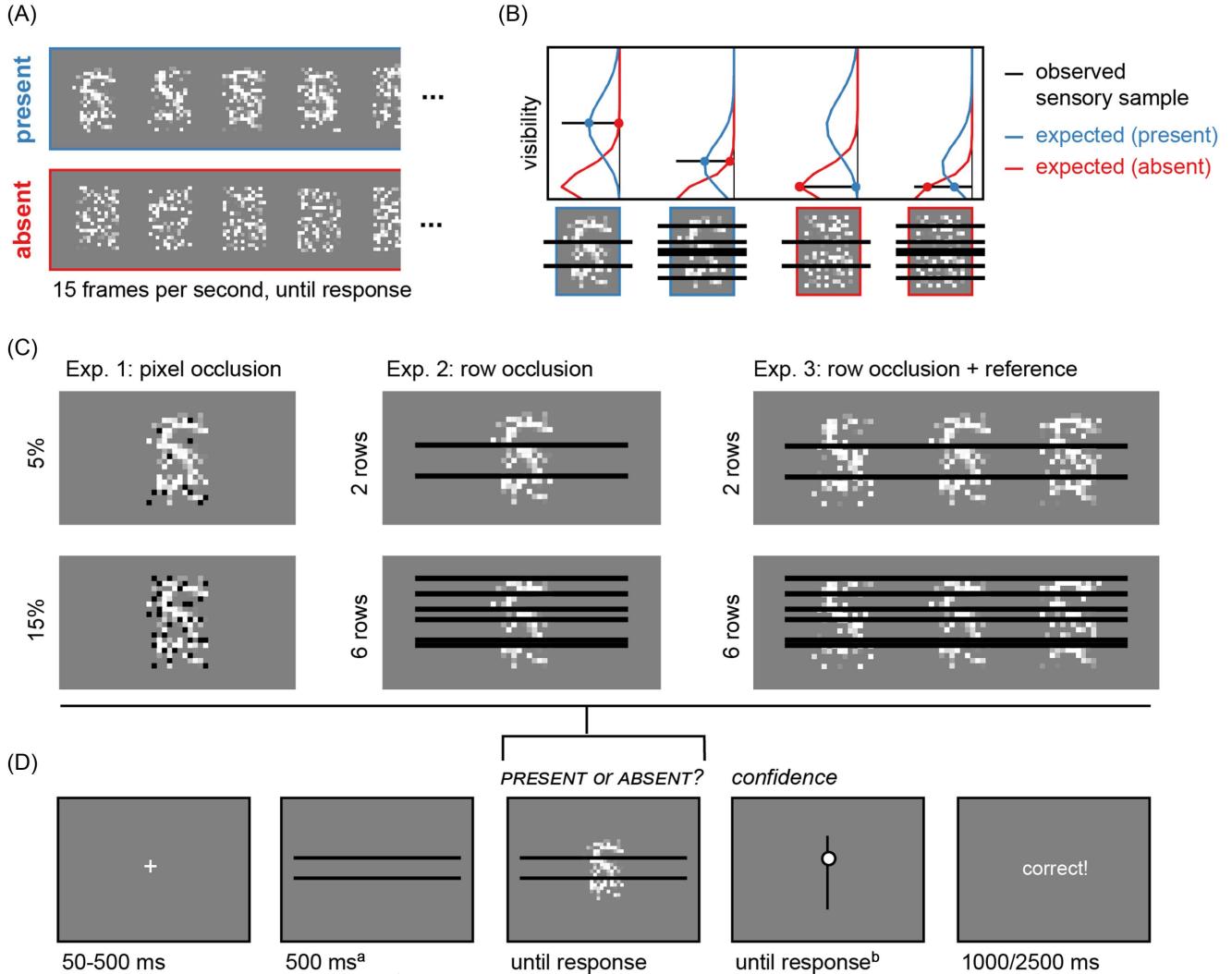
Crucially, agents can vary in the degree to which they flexibly incorporate such beliefs about visibility into their perceptual decisions. At one end of the spectrum, observers may “invert optics” using fully accurate beliefs about the evidence-generating process. At the opposite end, observers may be entirely unaffected by the expected likelihood of evidence, interpreting evidence at face value. Critically, while the former requires Bayesian inference, the latter strategy can be implemented as a rule-based process, by setting a criterion in sensory units (“respond absent for anything below this level of brightness,” Treisman & Williams, 1984) or temporal units (“respond absent if no sensory evidence for presence has become available within this much time,” Chun & Wolfe, 1996), without a reference to probabilities or likelihoods.

To ask where human observers fall on the continuum between these two opposite poles, we conducted three experiments where detection decisions about presence and absence were made under different levels of stimulus occlusion. Our task design allowed us to experimentally dissociate visibility (manipulated with occlusion and random fluctuations in the appearance of visual stimuli) from expected visibility (manipulated with occlusion only), independently measuring the effects of each on perceptual decisions and confidence in presence and absence. We show that behavioral asymmetries in perceptual detection naturally emerge in an ideal observer model when only presence, but not absence, is positively represented in sensory channels. Using the same ideal-observer model, we show that detection decisions in the absence of a stimulus—their biases, timing, and confidence—are critically dependent on beliefs about the expected visibility of stimuli that are not physically present, broadly consistent with “inverse optics” accounts of vision. Finally, qualitatively opposite behavioral effects in decisions about absence reveal reliable population heterogeneity in the incorporation of beliefs about visibility into perceptual decisions, independent of variability in visibility itself. We discuss the implications of possessing such beliefs about visibility, as well as these qualitatively different patterns across individuals, for perception and cognition more broadly.

## Method

### Task: Visual Detection Under Partial Occlusion

In three preregistered online experiments, participants performed a near-threshold detection task in which they made decisions about the presence or absence of a target letter (A or S, in different blocks) in a noisy, dynamic stimulus (Figure 2A). A target was present on a random 50% of trials. The stimulus remained on the screen, refreshing at 15 frames per second, until a response was made. On

**Figure 2***Rationale and Experimental Design for Experiments 1–3*

*Note.* (A) Example frames from target-present (blue) and target-absent (red) unoccluded stimuli. (B) Occluding more of a target letter decreases its visibility (measured, e.g., in sensory firing rates; black lines). Occlusion has a minimal effect on target visibility when the target is absent. Still, Bayesian observers should be affected by occlusion even when a target is absent due to its effect on expected visibility (blue and red curves), which is used to compare the likelihood of observed sensory samples under alternative world states (blue and red dots). Specifically, the likelihood ratio favors target absence much more strongly under low occlusion, due to an increase in the likelihood of the sensory observation under target presence rather than a change in the sensory observation itself. (C) Occlusion conditions in the three experiments. In Experiment 1, on different trials, we occluded a random subset of 5% or 15% of the pixels in the stimulus. In Experiments 2 and 3, on different trials we occluded a random subset of two or six pixel rows. In Experiment 3, the task-relevant stimulus was flanked by two reference stimuli that, known to the subject, always had the target letter in them. Participants performed two 16-trial blocks in which the target was the letter S and two blocks in which the target was the letter A. The order of the two letters was randomized between participants. (D) Trial structure in Experiment 2. Exp. = experiment. See the online article for the color version of this figure.

<sup>a</sup>The occluder preview screen only appeared in Experiments 2 and 3. <sup>b</sup>Confidence ratings were given only in Experiment 2, Blocks 3 and 4.

different trials, random parts of the display were occluded by an overlaid layer of static black occluders. Participants' task was to "ignore the black stuff, focus on the noise that is under it, and determine whether the letter appeared in it or not." We chose to manipulate stimulus visibility in this way, because the effect of occlusion on visibility is relatively obvious: The more occluded objects are, the harder they are to see. At the same time, the extent to which partial occlusion is expected to affect visibility depends on

observers' internal models of optics and of their visual systems. This way, we assumed that occlusion may affect not only stimulus visibility, but also beliefs about visibility, in turn affecting perceptual decisions even when a target is absent (Figure 2B). Further degrading stimulus visibility with dynamic visual noise allowed us to independently measure the effects of unmodelled variability in stimulus appearance on detection decisions. Since random fluctuations in stimulus appearance are unpredictable and

cannot be predicted by participants' internal models, we reasoned that they should affect decisions when a target is present much more than when it is absent.

More specifically, 253 English-speaking participants, recruited from the Prolific online platform, took part in Experiment 1, in which either 5% or 15% of the stimulus pixels were occluded by a static layer of randomly positioned black pixels (Figure 2C, left panels). 252 participants took part in Experiment 2, in which we occluded two or six entire rows of pixels, and presented the occluders for an additional 500 ms before stimulus onset, to facilitate a separation between the occluders and the noisy stimulus itself (Figure 2C, middle panels). Experiment 2 was the only experiment in which participants also reported their confidence ratings on an analog scale in Blocks 3 and 4 (Figure 2D). We decided not to include confidence ratings in the first two blocks so as not to contaminate the decision process with parallel processes relating to confidence ratings. Finally, 260 participants took part in Experiment 3, in which the central stimulus was flanked by two stimuli, partly hidden behind the same row occluders, which, known to participants, always had the target in them (Figure 2C, right panels). The rationale for this manipulation was to increase the availability of expected visibility ("what would a target look like?"), specifically in target-absent trials, making it easier for participants to reason "I would have seen the target if it were present."

After applying our preregistered accuracy and reaction-time-based exclusion criteria, 251, 234, and 250 participants were included in the main analysis for Experiments 1, 2, and 3, respectively. Results from all preregistered analyses are presented in the Appendix.

## Reverse Correlation Analysis

Since luminance values were randomly sampled per pixel and frame, the perceived similarity between the presented stimulus and the target letter fluctuated both within and between trials. This allowed us to directly measure the sensitivity of reaction times (RTs) in decisions about presence and absence to random fluctuations in stimulus–target similarity, quantified as the Pearson correlation between unoccluded pixels and their corresponding pixels in the target letter, statistically controlling for the overall effects of target presence, occlusion level, and calibrated visibility. Pearson's correlation was used since pixel luminance values were generated artificially and had no outliers.

For each individual frame, we computed the correlation between pixel luminance values and the corresponding values in the target image. Occluded pixels were omitted from both stimulus and target image. This resulted in an uncorrected stimulus–target similarity metric, which was naturally higher for frames from target-present trials, and for frames with higher levels of stimulus visibility ( $p$ , see the Procedure subsection). We therefore mean-centered these correlation values as a function of stimulus presence, occlusion and  $p$ , to obtain an unbiased measure of stimulus–target similarity  $r$ .

For each individual subject, we averaged these frame-wise correlations across frames 1–5, roughly corresponding to the first 300 ms of the trial (Mazor et al., 2023; Zylberberg et al., 2012). We then calculated, for each participant, the Spearman correlation between these average values and trial-wise RTs. This was done separately for target-present and target-absent trials. These subject-level correlations were then subjected to a group-level  $t$  test.

In Experiment 2, a similar procedure was used to reveal correlations between stimulus–target similarity and subjective confidence. For each participant, we calculated the Spearman correlation between trial-wise values of  $r$ , averaged across frames 1–5, and trial-wise confidence ratings, separately for target-present and target-absent trials. We then subjected the resulting correlations to a group-level  $t$  test.

## Sign Consistency Analysis

To quantify the within-subject consistency in the effect of occlusion on RT, we used a nonparametric sign-consistency test (Yaron et al., 2023). Individual-level sign-consistency scores were obtained by randomly splitting the trials of single participants into two disjoint sets, and measuring the effect of occlusion on response times within each half. We considered as "success" cases for which the two halves showed the same qualitative effect: slowing down or speeding up in the high-occlusion relative to the low-occlusion condition. By repeating this procedure 500 times and measuring the probability of success, we obtained a single sign-consistency score per participant, ranging from 0 to 1, where 1 indicates perfect sign consistency (for every split of the data, both halves showed the same qualitative effect) and 0.5 indicates chance sign consistency (the probability of success is equal to the probability of failure).

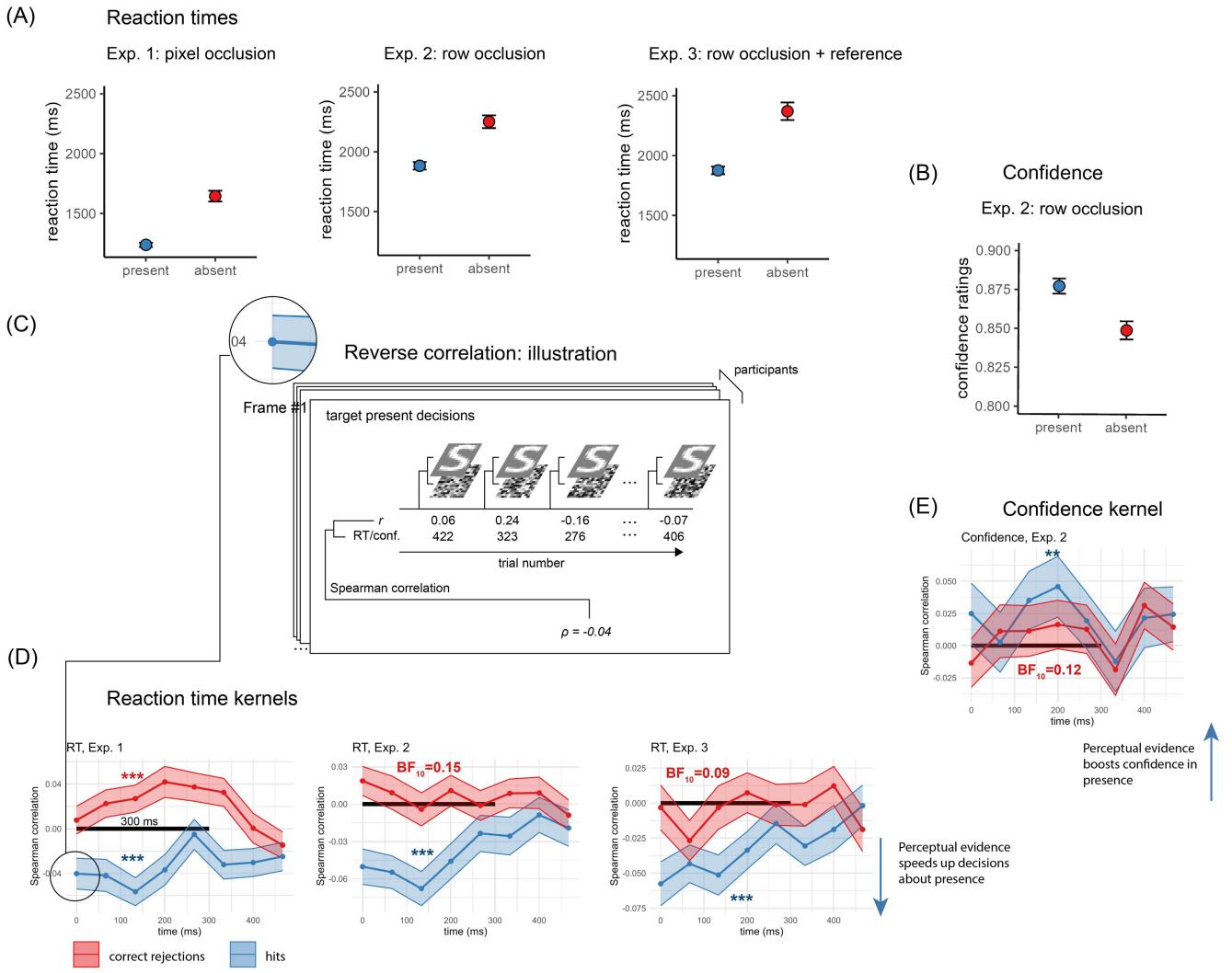
Averaging individual-level scores results in a group-level sign-consistency score. This score was compared to a null distribution, obtained by permuting the trial labels (high or low occlusion) before extracting 100 individual-level sign-consistency scores per participant, from which we extracted 10,000 group-level scores by averaging random subsets (Stelzer et al., 2013). A significant sign-consistency score indicates that the experimental manipulation had a consistent effect on individual participants.

## Results

### Presence–Absence Asymmetries

Before using decisions about absence as a critical test case for the role of beliefs about visibility in perceptual decisions, we need to establish that these decisions are made based on the absence of evidence for presence, and not based on direct evidence for absence. Crucially, if the absence of stimuli is perceived just like their presence, decisions about absence can be affected by occlusion simply because occlusion affects the visibility of evidence for absence (Gold & Shadlen, 2001), not because participants believe that it would affect the visibility of evidence for presence. Pronounced behavioral asymmetries between decisions about presence and absence provide initial, tentative evidence against a "direct perception of absence" account. First, correct decisions about absence were markedly slower than decisions about presence, by hundreds of milliseconds (Figure 3A). This was true in Experiment 1 (1.79 versus 1.33 s;  $t(250) = 17.40$ ,  $p < .001$ ), Experiment 2 (2.33 versus 1.95 s;  $t(233) = 15.90$ ,  $p < .001$ ), and Experiment 3 (2.49 versus 1.91 s;  $t(248) = 14.64$ ,  $p < .001$ ). Second, in Experiment 2, confidence ratings were significantly lower in decisions about absence (0.84 versus 0.86 on a 0.5–1 scale;  $t(233) = -2.92$ ,  $p = .004$ ; Figure 3B). While one could conceive of direct perception for absence, but poorer perception of

**Figure 3**  
Presence–Absence Asymmetries



*Note.* (A) Median reaction times in correct responses as a function of target presence. Error bars represent the standard error of the median, computed with bootstrapping. (B) Mean confidence in correct responses as a function of target presence. Error bars represent the standard error of the mean. (C) Analysis approach, illustrated for a single frame number. For each frame, the correlation between pixel luminance values and the target served as an index of stimulus–target similarity. This correlation was mean-centered separately as a function of occlusion level, target presence, and calibrated visibility levels. We then calculated, for each frame number and participant, the correlation between these similarity measures and their corresponding trial-wise reaction times or confidence ratings. This was done separately for correct target-present and target-absent decisions. Statistical tests were performed on average stimulus–target similarity values from the first 300 ms of the stimulus, represented by a black bar in panels D and E. (D) Mean correlations between confidence and stimulus–target similarity. Error margins are 1 standard error from the mean. (E) Mean correlations between confidence and stimulus–target similarity. RT = reaction time; conf. = confidence; Exp. = experiment; BF = Bayes factor. See the online article for the color version of this figure.

\*\*  $p < .01$ . \*\*\*  $p < .001$ .

it, these data at minimum suggest that evidence accumulation for presence and absence is not a symmetric process.

Furthermore, exploratory reverse correlation analysis revealed that RT and confidence ratings were driven by different factors in decisions about presence versus absence. Following previous reverse correlation studies of decision confidence (Mazor et al., 2023; Zylberberg et al., 2012), we focused our analysis on the first 300 ms of the stimulus presentation, and calculated, per trial, the mean similarity between the display and the target letter in these first

frames. We then computed the Spearman correlation between these trial-wise similarity measures and the RTs, focusing our analysis on correct responses only (see Figure 3C). Our reasoning was as follows: If perceptual evidence for both presence and absence equally contributes to perceptual detection decisions, effects of stimulus–target similarity on target-present RT and confidence should be mirrored by perfectly opposite effects of stimulus–target similarity in target-absent trials. If, however, only perceptual evidence for presence is represented, the negative effects of stimulus–target dissimilarity on

target-absent decisions should be attenuated relative to the corresponding effects of stimulus-target similarity on target-present decisions.

As expected, higher levels of stimulus-target similarity made participants quicker to detect the target letter when it was present, and this was the case in all experiments (a one-sample  $t$  test on within-subject correlation coefficients, calculated separately for the two occlusion levels and averaged per participant, Experiment 1:  $t(246) = -7.16, p < .001$ ; Experiment 2:  $t(250) = -7.72, p < .001$ ; Experiment 3:  $t(216) = -6.04, p < .001$ ; blue curves in Figure 3D). In contrast, higher levels of stimulus-target similarity made participants slower to notice the absence of the letter when it was absent only in Experiment 1 ( $t(246) = 4.57, p < .001$ ), but not in Experiment 2 ( $t(250) = 1.22, p = .222, BF_{10} = 0.15$ ), and Experiment 3 ( $t(218) = -0.32, p = .748, BF_{10} = 7.79 \times 10^{-2}$ ). In all cases, the effect of stimulus-target similarity on decision times was stronger in target-present compared to target-absent responses (Experiment 1:  $t(246) = -2.07, p = .039$ , Experiment 2:  $t(249) = -4.59, p < .001$ , Experiment 3:  $t(216) = -4.77, p < .001$ ; red curves in Figure 3D).

Confidence ratings in Experiment 2 further allowed us to test the relationship between stimulus-target similarity and subjective confidence, revealing a similar pattern. Confidence judgments in hits were positively correlated with stimulus-target similarity in the first 300 ms of the trial ( $t(234) = 3.15, p = .002$ ; blue curve in Figure 3E). In contrast, confidence in correct identifications of target absence showed no negative relationship to perceptual evidence ( $t(242) = 0.81, p = .418, BF_{10} = 0.12$ ; red curve in Figure 3E). Similar to RTs, the difference between the effect of perceptual

evidence on confidence in presence and the (negative) effect of perceptual evidence on confidence in absence was in itself significant ( $t(232) = 3.14, p = .002$ ). Unlike confidence in presence, confidence in absence was not based on dissimilarity to the target letter.

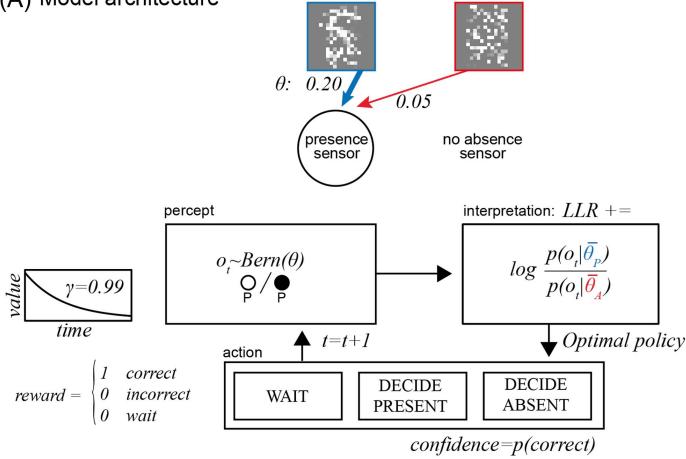
## An Ideal Observer Model of Visual Detection

Presence-absence asymmetries in RT and confidence are expected if evidence is only ever available to support presence, leaving absence to be inferred tentatively and based on the absence of evidence. To formulate this asymmetry in the availability of evidence, we present a partially observed Markov decision process (POMDP; Littman, 2009) model of perceptual detection. Critically, our model has an asymmetric structure: It is equipped only with a presence-sensor, but not with an absence-sensor (Figure 4A, right panels). The sensor produces sequences of activations and inactivations according to a true activation probability (similar to a neuronal firing rate), which is sensitive to the presence of the target and to its visibility. We ask whether, when faced with this evidence structure, a rational agent would behave in ways that resemble the behavior of our participants (Anderson, 1990). We provide a high-level description of the model here and a more detailed description in the Extended Method section. As we show, presence-absence asymmetries in decision time and decision confidence are borne out of rational evidence accumulation when the information value of evidence for presence and absence is itself asymmetrical.

We model sensory observations as the binary (on/off) activations of a “presence sensor” which is probabilistically tuned to one state of the world. For example, in our illustration (Figure 4A), the presence

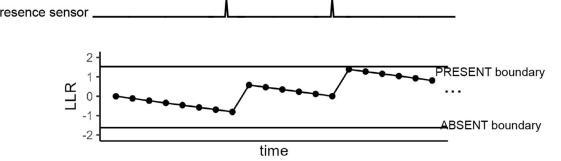
**Figure 4**  
*An Ideal Observer Model of Visual Detection*

(A) Model architecture

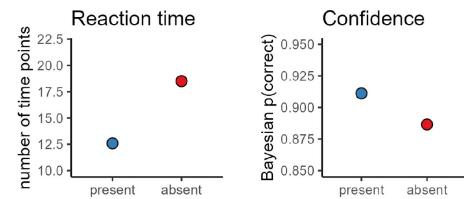


**Note.** (A) Model architecture. Target presence affects the activation probability of a presence sensor. The agent perceives a series of binary outcomes (sensor activations), based on which it attempts to guess the true world state (target presence or absence). This is done by extracting and accumulating the log likelihood ratio (LLR) for target presence versus absence and deciding whether to make a decision based on available evidence, or alternatively whether to accumulate more evidence in order to obtain a better estimate of the true activation probability. In making this decision, the agent balances the incentive to be as accurate as possible (only correct decisions are associated with an intrinsic reward) and the exponential discounting of the value of reward as a function of time. (B) The first samples from an example trial and their interpretation by the agent. Sensor activations are much more informative than sensor inactivations, as indicated by the shallow negative and steep positive slopes of the LLR sequence. (C) Behavioral predictions for an ideal observer. The model predicts that decisions about absence should be slower, and that they should be accompanied by lower levels of subjective confidence, than decisions about presence. See the online article for the color version of this figure.

(B) Example trial



(C) Predictions for an ideal observer:



sensor has a higher activation probability when a target is present (0.20) than absent (0.05). The agent is intrinsically rewarded for making accurate decisions regarding the state of the world (stimulus presence or absence), given these noisy observations. To increase the probability of being correct, the agent can choose to wait and accumulate more observations before making a decision. However, the intrinsic value of accuracy is subject to exponential temporal discounting, rendering the value of later correct decisions lower than that of earlier ones. Thus, from the agent's perspective, accumulating further evidence pays off exactly when the expected accuracy gain exceeds the discounting loss.

Given these settings, our agent implements an optimal policy, updating its beliefs about the state of the world by tracking the log likelihood ratio (LLR) between presence and absence given each observation, and committing to a decision only when the expected value of making a decision now is higher than the expected value of making a decision later, assuming the same policy will be used in later time points. We obtained the optimal policy—a probabilistic mapping from belief states to perceptual decisions or further evidence accumulation—using backward induction (Callaway et al., 2024; Puterman, 2014; Tajima et al., 2016). Although we did not explicitly assume a priori that LLR was integrated to a boundary, the resulting optimal policy can still be described in these terms. Optimal choices are made when the LLR either exceeds a positive boundary, indicating sufficient evidence for presence, or a separate negative boundary, indicating sufficient evidence for absence (note these two boundaries are not necessarily symmetric around 0). Crucially, these boundaries are implied by the optimal policy and are not themselves model parameters. As an additional measure, we assume that decision confidence equals the probability of being correct at the time of committing to a decision, given the accumulated evidence so far.

Since sensor inactivation is the more likely state both when a stimulus is present and absent, it is much less informative than sensor activation: for the parameters used here (sensor activation probabilities of 0.20 and 0.05 when a target is present or absent, respectively), activation is four times more likely when a target is present than absent, but inactivation is more likely by a factor of only 1.18, when a target is absent than present (see the smaller steps going down compared to up in Figure 4B). This asymmetry in the information value of evidence for presence and absence results in further asymmetries in the position of the implied decision boundaries (the upper boundary is slightly closer to the midpoint in Figure 4B). Together, this produces slower decisions about absence and lower confidence levels in such decisions compared to decisions about presence (Figure 4C). These presence–absence asymmetries are consistent with the RT and confidence profiles found in perceptual detection experiments (Kellij et al., 2021; Mazor et al., 2020, 2021, 2023; Meuwese et al., 2014).

This model makes several simplifying assumptions that are important to acknowledge. First, the probability of sensor activation is assumed to be a function of target presence only, invariant to systematic changes in evidence weighting as a function of time-point within a trial. Our model therefore does not explicitly model the increased sensitivity to visual information in the first 300 ms of stimulus presentation (Mazor et al., 2023; Zylberberg et al., 2012). Second, temporal discounting is assumed to take an exponential form into an infinite temporal horizon, while impulsive observers may disproportionately value the present (Ainslie, 2017). Finally, confidence is assumed to be based on the decision variable itself,

with no postdecisional evidence accumulation (Moran et al., 2015; Petrusic & Baranski, 2003; Yeung & Summerfield, 2012) and is assumed to be identical to the Bayesian probability of being correct at the time of making the decision. These simplifying assumptions enabled us to efficiently obtain optimal policies for different sets of beliefs and preferences as part of the model fitting process and without introducing additional free variables.

### **Modeling Occlusion Effects**

In light of these marked presence–absence asymmetries in RTs, confidence ratings, and evidence weighting, we proceed with the assumption that positive evidence for absence is not explicitly coded in sensory channels. We simulate stimulus occlusion as a scaling of the probability of sensor activation by a parameter  $\alpha \in [0,1]$ , such that  $p(1|0,\alpha) = \alpha 0$ . This way,  $\alpha$  can be thought of as modulating the visibility of target-like patterns, with lower levels making the sensor less likely to activate (Figure 5A). Importantly, in addition to the effects of  $\alpha$  on stimulus visibility, beliefs about  $\alpha$  (denoted  $\bar{\alpha}$ ) also affect how sensory input is interpreted, and how much certainty agents seek before they commit to a decision. Figure 4B illustrates the interpretation of the same sensory samples, when the agent believes  $\alpha$  to be .8 (corresponding to low occlusion, in black) or .6 (corresponding to high occlusion, in gray). Notably, the information value (measured as  $|LLR|$ ) of sensor inactivation, but not sensor activation, is diminished when  $\bar{\alpha} = .6$ , making the same ambiguous sequence of samples appear more consistent with target presence if the display is known to be occluded.

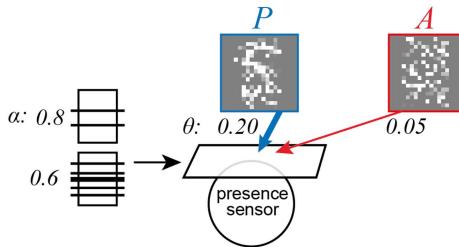
In this model, occlusion affects the probability of obtaining positive evidence, but beliefs about occlusion have no effect on the interpretation of such evidence once obtained. This is true because while the overall probability of obtaining positive evidence  $p(1)$  diminishes with higher levels of occlusion, the *relative* probability of such evidence given target presence or absence  $\frac{p(1|\text{present})}{p(1|\text{absent})}$  remains unaffected. On the other hand, occlusion has little effect on the probability of obtaining negative evidence (in the form of sensor inactivation), but beliefs about the effects of occlusion affect the interpretation of such evidence once obtained. As a result, the timing and confidence of perceptual decisions in the absence of a target depend much more on beliefs about the effect of occlusion than on the true effect of occlusion on visibility.

To exemplify the dissociable contributions of visibility itself ( $\alpha$ ) and beliefs about visibility ( $\bar{\alpha}$ ) to behavior, we consider two variants of the model (Figure 5B). Variant  $V_{\text{IGNORE}}$  entirely ignores the expected effects of  $\alpha$  on the probability of sensor activation, interpreting evidence in the same way in both high-occlusion and low-occlusion trials. This variant serves as our null model, in that it specifies that occlusion affects perception only to the extent that it affects the activation of sensors, but not in how these activations are interpreted or accumulated over time. Crucially, this model variant can be implemented without reference to the perceptual likelihood function, by specifying weights on sensor activation and inactivation, and fixed decision thresholds.

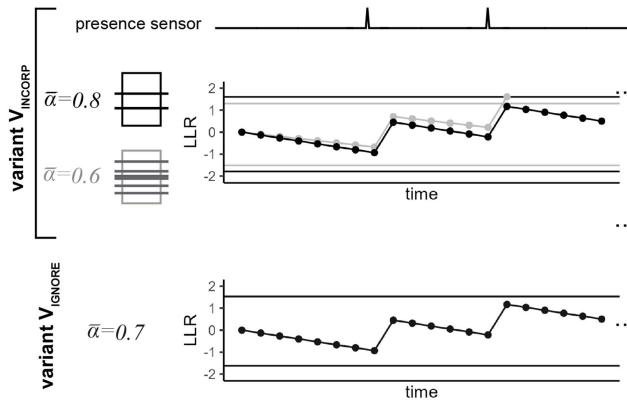
$V_{\text{INCORP}}$ , on the other hand, incorporates into its perceptual decisions fully accurate beliefs about the effect of  $\alpha$  on stimulus visibility. This affects both the interpretation stage (when  $\alpha$  is believed to be low, sensor inactivation, but not sensor activation, becomes less informative) and the action selection stage (by affecting the expected value of future evidence, making the agent more willing

**Figure 5**  
Modeling the Effects of Occlusion on Visual Detection

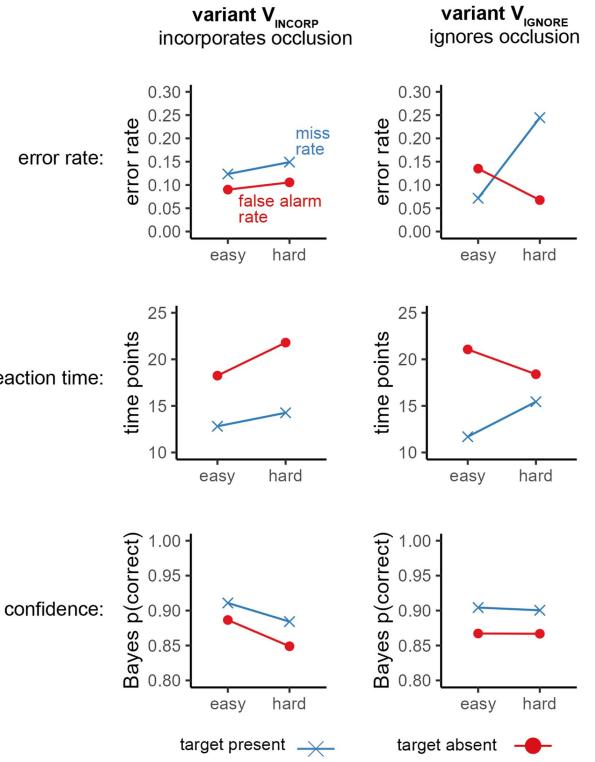
(A) Model architecture



(B) Example trials



(C) Model predictions



**Note.** (A) We model the effect of occlusion as scaling of the probability of sensor activation. (B) The first sensory samples from example trials and their corresponding interpretations as a function of the believed visibility level. (C) Reaction time and confidence effects in simulated rational observers as a function of target presence and absence and level of occlusion, correct trials only. Variant  $V_{\text{INCORP}}$  considers the effect of occlusion when interpreting sensory evidence and making decisions, whereas variant  $V_{\text{IGNORE}}$  does not. LLR = log likelihood ratio. See the online article for the color version of this figure.

to settle for lower decision confidence when occlusion is high). Importantly, variant  $V_{\text{INCORP}}$  represents one point in the space of possible values  $\bar{\alpha}$  can take relative to  $\alpha$ : Observers who incorporate beliefs about visibility into their perceptual decisions may underestimate the effect of occlusion on visibility or overestimate it.

The two model variants predict different effects of occlusion on accuracy, decision times, and confidence ratings. This is especially evident in target-absent trials (red lines in Figure 5C). While occluding more of the display makes  $V_{\text{INCORP}}$  commit more false alarms, it makes  $V_{\text{IGNORE}}$  make fewer of them.  $V_{\text{INCORP}}$ 's decisions about absence are slower when more of the display is occluded, whereas  $V_{\text{IGNORE}}$ 's decisions about absence are faster. Finally,  $V_{\text{INCORP}}$  is less confident in decisions about absence when more of the display is occluded, but this is not true of  $V_{\text{IGNORE}}$ . Together, both the size and direction of occlusion effects on decisions about absence are dependent on meta-perceptual knowledge about the influence of occlusion on visibility or the incorporation of such knowledge into perceptual decisions.

### Occlusion Effects

Equipped with the predictions of the two model variants, we now turn to the experimental data.

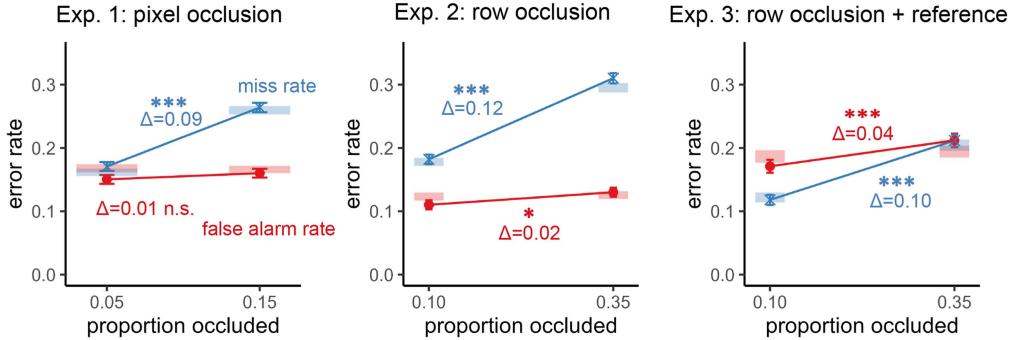
### Target-Present Trials

In all three experiments, occlusion had the expected effects on the detection of present targets (see Figure 6A). Specifically, participants missed more targets when more of the display was occluded. In Experiment 1, the mean hit rate went down from 0.81 when 5% of the pixels were occluded to 0.72 when 15% of the pixels were occluded ( $t(250) = 8.92, p < .001$ ; see Figure 6A, blue lines). In Experiment 2, the mean hit rate went down from 0.78 when two rows of pixels were occluded to 0.66 when six rows were occluded ( $t(233) = 11.83, p < .001$ ), and similar figures were obtained in Experiment 3 (from 0.85 to 0.75 in the two occlusion levels;  $t(248) = 12.71, p < .001$ ).

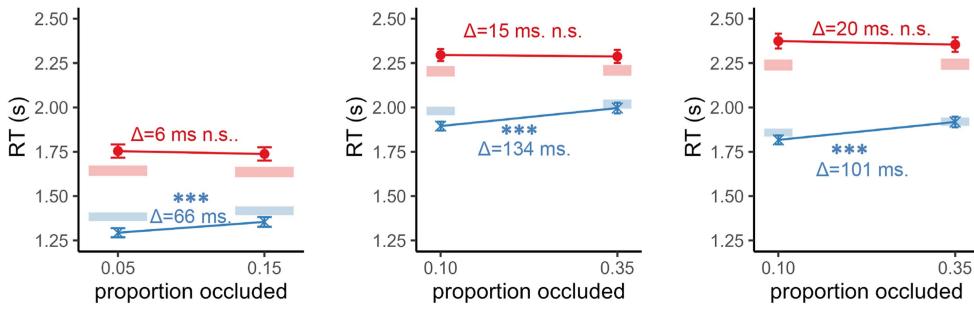
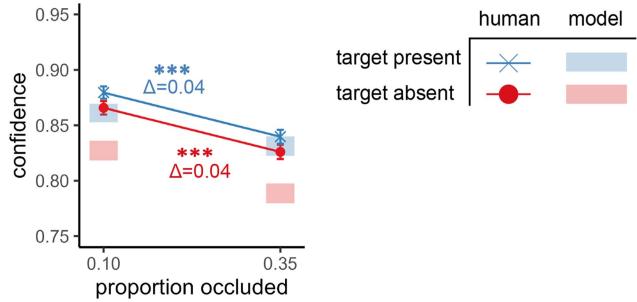
Participants were also slower to correctly detect targets when more of the display was occluded, with a mean difference of 66 ms in Experiment 1 ( $t(250) = -3.51, p < .001$ ), 134 ms in Experiment 2 ( $t(233) = -5.13, p < .001$ ), and 101 ms in Experiment 3 ( $t(248) = -5.19, p < .001$ ; see Figure 6B, blue lines). In all three experiments, this effect remained significant when incorporating incorrect trials into the analysis. Finally, in Experiment 2, confidence in presence was lower when more of the display was occluded (0.84 versus 0.88 on a 0.5–1 scale;  $t(233) = 9.87, p < .001$ ; see Figure 6C, blue line).

**Figure 6**  
Main Results From Experiments 1–3

## (A) error rates



## (B) reaction times

(C) confidence  
(out of sample prediction)

*Note.* (A) Miss and false alarm rates as a function of occlusion level. (B) Mean median reaction times in target-present and target-absent correct responses, as a function of occlusion level. (C) Mean confidence in target-present and target-absent correct responses, as a function of occlusion level. Error bars represent the standard error. Semitransparent rectangles represent data simulated from the model, fitted to accuracy, and reaction time (but not confidence) data of individual participants (see the Model fitting section). Rectangles are centered at the mean value, and their height is twice the standard error. Exp. = experiment; RT = reaction time; n. s. = not statistically significant. See the online article for the color version of this figure.

\*  $p < .05$ . \*\*\*  $p < .001$ .

Unsurprisingly, occluding more of the target made it more difficult to spot.

**Target-Absent Trials**

Having established that occlusion affected stimulus visibility, making responses slower, less accurate, and accompanied by lower

levels of confidence when a target was present, we next examined the effects of occlusion on detection responses in the absence of a target. If participants were, like model variant  $V_{INCP}$ , effectively incorporating beliefs about visibility into their criterion placement, we would expect to see an increase in the proportion of false alarms (the proportion of incorrect target-present reports out of all target-absent trials), when more of the target was occluded. If, however,

they took evidence at face value like model variant  $V_{\text{IGNORE}}$ , occlusion should be expected to reduce the false alarm rate.

We did not observe an effect of occlusion on the false-alarm rate in Experiment 1 (equaling 0.17 and 0.18 when 5% or 15% of the stimulus pixels were occluded,  $t(250) = -1.16, p = .249$ ). We therefore made the occlusion manipulation clearer to participants in Experiments 2 and 3, occluding entire rows and up to 35% of the display. An increase in the false alarm rate with higher levels of occlusion, consistent with variant  $V_{\text{INCORP}}$ , was observed in Experiment 2 (from 0.13 to 0.15 when two or six rows were occluded,  $t(233) = -2.26, p = .025$ ), and 3 (from 0.19 to 0.23 when two or six rows were occluded,  $t(248) = -4.98, p < .001$ ). A significant increase in the effect of occlusion on the false alarm rates between Experiments 2 and 3 (a between-subject  $t$  test,  $t(480.40) = -2.09, p = .037$ ) is consistent with the use of beliefs about visibility to make decisions in the absence of a target: in Experiment 3, but not in Experiment 2, the central stimulus was flanked by two target-present stimuli which were hidden behind the same occluders, making the effect of occlusion on visibility visually available even in target-absent trials (see Figure 2C, right panel).

In line with the predictions for variant  $V_{\text{INCORP}}$ , but not  $V_{\text{IGNORE}}$ , confidence in absence was lower when more of the display was occluded (0.83 versus 0.87;  $t(233) = 10.54, p < .001$ ). Furthermore, we find no difference between the effects of occlusion on confidence as a function of target presence ( $t(233) = 0.01, p = .992$ ,  $\text{BF}_{10} = 7.32 \times 10^{-2}$ ; see Figure 6C). Participants were less confident in the absence of a target when it would have been harder to see.

Finally, the two model variants made opposite predictions for the effect of occlusion on RTs in the absence of a target. While model variant  $V_{\text{INCORP}}$  predicted slower decisions about absence with more occlusion, model variant  $V_{\text{IGNORE}}$  predicted the opposite pattern. Intriguingly, an effect of occlusion on target-absent RTs did not emerge in any of the three experiments. Specifically, we observed a mean difference of 6 ms in Experiment 1 ( $t(250) = 0.30, p = .765$ ;  $\text{BF}_{10} = 7.39 \times 10^{-2}$ ), 15 ms in Experiment 2 ( $t(233) = 0.79, p = .429$ ;  $\text{BF}_{10} = 9.97 \times 10^{-2}$ ), and 20 ms in Experiment 3 ( $t(248) = 0.68, p = .498$ ;  $\text{BF}_{10} = 9.97 \times 10^{-2}$ ;  $\text{BF}_{10} = 6.91 \times 10^{-2}$  when pooling data from all three experiments), reflecting strong evidence for the null hypothesis (see Figure 6B, red lines).

## Additional Data Reveals Individual Differences in Occlusion Effects on Inference About Absence

Occlusion affected the false alarm rate and subjective confidence in a way that is consistent with the incorporation of counterfactual visibility into inferences about absence, but the absence of an effect on decision time was inconsistent with both models: Model  $V_{\text{INCORP}}$  predicted a positive effect and model  $V_{\text{IGNORE}}$  a negative one. We considered the possibility that this null group-level result may reflect population variability in the incorporation of beliefs about visibility into perceptual decisions, with some behaving more in line with the prediction of model  $V_{\text{INCORP}}$ , incorporating expected visibility into their perceptual decisions about absence and slowing down when more of the display is occluded, and others more in line with the predictions of model  $V_{\text{IGNORE}}$ , underestimating the effect of occlusion on stimulus visibility or ignoring it altogether, resulting in speedier decisions about absence for more occluded displays.

This population-mixture model predicts that despite a group-level null effect, some individual participants should show reliable effects of occlusion on “target absent” RTs: negative for some participants, and positive for others. To test this prediction, we collected a large number of test trials from a random subset of ten participants who took part in Experiments 2 and 3 (see Figure 7A). Over the course of five sessions we collected 896 trials per participant, with the exception of two participants in Experiment 3 for which we have 672 and 864 trials.

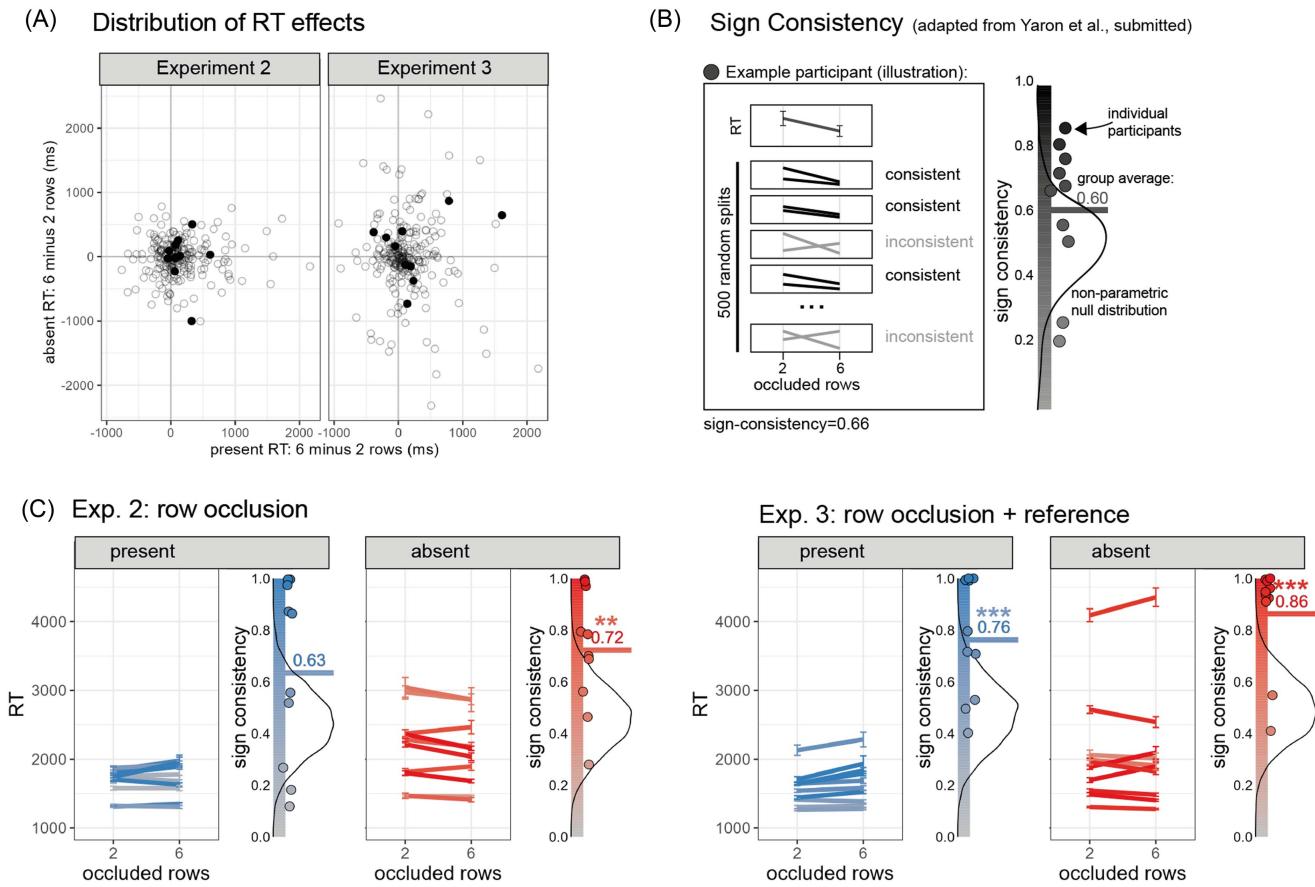
The high number of trials per participant allowed us to quantify the consistency of the effect of occlusion on target-absent RTs within individual participants. For each participant, we compared their target-absent response times in high- and low-occlusion trials with a  $t$  test. If decision times were invariant to the effect of stimulus occlusion, this would be expected to result in a significant test statistic in one of 20 participants, on average, corresponding to our significance level of 0.05. Strikingly, however, out of 20 participants, the effect of occlusion on “target absent” decision times was significant in eight, split exactly half-half between significant positive effects (more consistent with model variant  $V_{\text{INCORP}}$ ) and significant negative effects (more consistent with model variant  $V_{\text{IGNORE}}$ ): much higher than the 1/20 probability expected by chance alone ( $p < .001$  in a binomial test against  $p = .05$ ).

As a more sensitive test of effect reliability, we employed the nonparametric *sign-consistency test* (Yaron et al., 2023): randomly splitting individual participants’ trials into two subsets, and asking whether both subsets demonstrate the same type of outcome: either positive or negative (see Figure 7B and the Method section). The group-level mean sign consistency, or the proportion of these random splits where the same outcome is observed in both subsets, is then compared against a bootstrapped null distribution to obtain a group-level  $p$  value.

In both experiments we find clear evidence for above-chance sign consistency in the effects of occlusion on RTs in target-absent trials (Experiment 2: sign consistency = 0.72,  $p = .003$ ; Experiment 3: sign consistency = 0.86,  $p < .001$ ; see Figure 7C). Moreover, target-absent sign-consistency scores were not significantly different from, and numerically higher than, target-present sign-consistency scores (Experiment 2: sign consistency = 0.63,  $p = .073$ ; Experiment 3: sign consistency = 0.76,  $p < .001$ ). These data provide clear evidence that an effect of expected visibility on “target absent” response times was not absent in Experiments 2 and 3: It was masked by differences between individual participants who systematically exhibit positive and negative effects.

Interestingly, this population variability in target-absent response times aligned with population variability in the effect of occlusion on the false alarm rates. In both long experiments, participants who were slower to make decisions about absence in the high-occlusion condition were also more likely to make more false alarms in this condition (Spearman’s correlation between occlusion effects on target-absent RTs and on the false alarm rate:  $r_s = .47, S = 704.00, p = .038$ ). This positive correlation is the opposite of what is expected based on a speed–accuracy trade-off, which would predict an increase in accuracy among participants who slow down in the high-occlusion condition. In contrast, it is expected if participants varied in the degree to which expected visibility fed into their perceptual decisions, that is, in their alignment with  $V_{\text{INCORP}}$  versus  $V_{\text{IGNORE}}$ .

**Figure 7**  
Sign Consistency Analysis



*Note.* (A) Occlusion effect distributions in Experiments 2 and 3. Individual participants are represented as circles. In order to obtain sufficient statistical power, we collected hundreds of trials from a random subset of 10 participants who participated in each experiment (marked with filled circles). (B) An illustration of the sign-consistency test, for a hypothetical participant. Sign consistency is the proportion of random splits, out of 500, for which both trial subsets show the same qualitative effect. Individual sign-consistency scores are then averaged and compared against a nonparametric null distribution to obtain a *p* value. (C) Sign-consistency results. Within each panel, we present median RT as a function of occlusion level for each participant on the left. Color saturation indicates sign consistency. On the right, we present individual sign-consistency scores as circles, alongside the group-average sign consistency score (horizontal line), overlaid on top of the nonparametric null distribution. In both experiments, group-level sign consistency was significantly above chance for the effect of occlusion on response time in target-absent trials. RT = reaction time; Exp. = experiment. See the online article for the color version of this figure.

\*\* *p* < .01. \*\*\* *p* < .001.

### Model Fitting Reveals Individual Differences in the Incorporation of Expected Visibility Into Perceptual Decisions

In order to map this behavioral variability onto the model parameter space, we fitted model parameters to the response and response time data of participants (but not to confidence ratings, which were treated as an out-of-sample test for our model predictions; see Extended Method section for details about the model fit procedure and Appendix for parameter and model recovery results).

Specifically, we fitted four model variants to response and response time data. The full “inverse optics” model had an asymmetric architecture, with a presence sensor but no absence sensor. In this

model, the physical visibility of stimuli (parameters  $\theta$  and  $\alpha$ ) and participants’ beliefs regarding these quantities (parameters  $\bar{\theta}$  and  $\bar{\alpha}$ ) were allowed to independently vary. As a result, this model could capture miscalibrated beliefs about the effect of occlusion on visibility, such as a belief that occlusion has an effect on visibility, but a lesser effect than its actual, true effect. A “symmetric” model was similar to the “inverse optics” model but had an absence sensor with a perfectly symmetric tuning curve to that of the presence sensor. Together, this produces a neuron/antineuron architecture, in which perceptual decisions can be made in a rule-based manner, and without reference to likelihoods (Gold & Shadlen, 2001). An “IGNORE” variant was similar to the “inverse optics” variant, with the exception that the agent was assumed not to incorporate beliefs

about the effect of occlusion on visibility into their decisions. Like the symmetric model, this variant too can be implemented as a fixed decision rule and without reference to likelihoods or probabilities. Finally, an “INCOP” variant was similar to the “inverse optics” variant, but agents were assumed to incorporate perfectly accurate beliefs about the effect of occlusion on visibility.

For a formal model comparison, we used data from the 20 participants from whom we collected hundreds of trials. We used the Akaike information criterion to identify the best fitting model for each individual participant, as this criterion showed good model recovery (see Appendix). Out of 20 participants, one participant was best fitted by the symmetric model, six were best fitted by the IGNORE model, four were best fitted by the INCOP model, and the remaining nine were best fitted by the full “inverse optics” model, in which visibility and beliefs about visibility varied independently. Given that IGNORE and INCOP are private cases of the “inverse optics” model, we continue with describing the fit of this more general model to the data.

As shown in Figure 6 (semitransparent rectangles), the model captures key aspects in the data, including the difference in RTs as a function of target presence or absence, the increase in “target present” RTs as a function of occlusion, and the invariance of “target absent” RTs to occlusion. Despite being fitted to response and response time data only, the model also captures the two main qualitative patterns in confidence ratings: lower confidence in absence than in presence judgments, and lower confidence in high versus low occlusion. Furthermore, perceptual evidence, quantified as sensor activation sequences, contributed more to decision time and decision confidence when a target was present than absent, mirroring the reverse correlation findings from human data (see Appendix). For comparison, the IGNORE variant fitted the data relatively well in target-present trials, predicting an increase in RT and in the error rate, but it made incorrect predictions about target-absent trials, predicting a decrease in RT and a reduction in the false alarm rate (see Appendix).

Notably, the full model captures the overall bias to report absence in Experiment 2 (0.56 of all observed and simulated responses) and presence in Experiment 3 (0.51 of all simulated responses, 0.52 of all observed responses), although the prior probability of target absence is assumed to be known to be 0.5 in both experiments, and the incentive structure is fully symmetric with respect to false alarms and misses. A response bias emerges due to asymmetries in the likelihood function and the different information value of evidence for presence versus absence.

Some aspects of the data were not captured by our model, including the group-level effect of occlusion on the false alarm rate in Experiments 2 and 3 and the relatively slower error trials (in the Appendix, we show that allowing visibility and beliefs about visibility to vary between trials can account for the relationship between accuracy and RT). Furthermore, while our model captured the fact that “target-absent” responses are overall slower, it underestimated this effect (see Figure 6B), suggesting that other processes, perhaps outside perception (Beltrán et al., 2021; Zang et al., 2022), contribute to the difference in response times between target-present and target-absent responses. However, this seems a valid assumption. Finally, while reverse correlation analysis of simulated sensor activations produced the qualitative asymmetry between evidence weighting leading to target-present or target-absent decisions, it did not produce the characteristic temporal

profile of reaction time and confidence kernels observed in human observers (Mazor et al., 2023; Zylberberg et al., 2018), suggesting that other factors contribute to observers’ increased perceptual sensitivity early in the trial.

The model successfully captured some, but not all, population variability in the effects of occlusion on error rates and RTs. Specifically, the Spearman correlation between the effect of occlusion on target-absent decision times in human data and in simulated data, generated using the parameters fitted to individual subjects, was  $r_s = .28$  in Experiment 1,  $r_s = .46$  in Experiment 2, and  $r_s = .43$  in Experiment 3.

Inspecting the fitted model parameters revealed that overall,  $\alpha$  and  $\bar{\alpha}$  were correlated across individuals (Experiment 1:  $r_s = .60$ ; Experiment 2:  $r_s = .55$ ; Experiment 3:  $r_s = .67$ ; see Appendix for full distributions), meaning that participants’ beliefs about the effects of occlusion on visibility were proportional to the true effect of occlusion on stimulus visibility. Importantly, despite this strong alignment, participants had an overall tendency to act in accordance with a belief that occlusion affected visibility to a lesser degree than its true effect (Experiment 1:  $t(250) = 5.56$ ,  $p < .001$ ; Experiment 2:  $t(233) = 6.72$ ,  $p < .001$ ; Experiment 3:  $t(248) = 5.04$ ,  $p < .001$ ). While occlusion had a similar effect on  $\alpha$  in Experiment 2 and 3 ( $t(475.56) = 0.59$ ,  $p = .552$ ), the added reference stimuli in Experiment 3 affected  $\bar{\alpha}$ , bringing it closer to  $\alpha$  itself (a contrast between  $|\alpha - \bar{\alpha}|$  in Experiments 2 and 3:  $t(461.16) = 2.14$ ,  $p = .033$ ). This is in line with the theoretical interpretation of these two model parameters as being based in the true and believed effects of occlusion on visibility, respectively.

Finally, we compared the effects of  $\alpha$  (true visibility) and  $\bar{\alpha}$  (beliefs about visibility) on accuracy and RTs in target-present and target-absent trials by extracting Spearman’s correlations between fitted parameters and individual-level contrasts of interest. A clear picture emerges: While  $\alpha$  explains more variance than  $\bar{\alpha}$  in decision accuracy and decision times in target-present trials (first two coloured columns in Table 1), the opposite is true for target-absent trials (last two coloured columns in Table 1). Together, model fit results confirm that beliefs about visibility play a role in perceptual decisions, that this is especially true in decisions about target absence, and that this is subject to significant population variability.

**Table 1**  
*Spearman’s Correlations Between Model Parameters and Occlusion Effects on Accuracy and Reaction Times*

Model parameter	Exp.	Miss	RT present	False alarm	RT absent
$\alpha$ (true visibility)	1	.69***	-.16**	-.048	.069
	2	.71***	-.21**	-.088	-.054
	3	.52***	-.17**	-.07	.074
$\bar{\alpha}$ (believed visibility)	1	.18**	-.009	.31***	-.052
	2	.13	-.046	.39***	-.19***
	3	.066	.021	.37***	-.1

*Note.* Darker shades of blue indicate lower  $p$  values. The visibility parameter is correlated with occlusion effects when a target is present (first two coloured columns), but the parameter controlling beliefs about visibility is correlated with occlusion effects when a target is absent (right two columns). Exp. = experiment; RT = reaction time. See the online article for the color version of this table.

\*\*  $p = .01$ . \*\*\*  $p = .001$ .

## Discussion

Much focus has been placed on the role of prior expectations in perceptual inference (Kok et al., 2013; Powers et al., 2017; Press et al., 2020; Summerfield & Egner, 2009; Weilnhammer et al., 2018; Yon et al., 2021, 2023), with important discussions regarding the (im)penetrability of visual perception to such effects from cognition (Firestone & Scholl, 2016; Pylyshyn, 1999) and potential implications for models of delusions and schizophrenia (Corlett et al., 2019; Haarsma et al., 2023; Powers et al., 2017; Stuke et al., 2019). Here, we focus on the other component of Bayesian reasoning, often neglected in such discussions: beliefs about the mapping from world states to sensory input. Unlike prior expectations about the world (e.g., the probability that someone would be knocking on my door), beliefs about the likelihood of observations given world states describe the perceptual system itself (e.g., the probability that I would be able to hear the knock if someone was knocking on my door). Previous work postulated a role for such meta-perceptual beliefs in implicit and explicit measures of decision confidence (Hellmann et al., 2023; Herce Castañón et al., 2019; Olawole-Scott & Yon, 2023; Rausch et al., 2018), and in dissociations between objective and subjective measures of awareness (Ko & Lau, 2012). Focusing on a visual detection task, and specifically on trials in which no target was present, we show that beliefs about visibility affect decision times and decision criteria of the detection judgments themselves.

Our novel ideal observer model of perceptual detection, formalized as a POMDP, successfully accounts for key signatures of perceptual detection, including an overall bias to report absence or presence, and systematic differences in the timing and confidence with which decisions about presence and absence are made. Critically, and unlike extant models of perceptual decision-making, this is done not by manipulating the observers' prior belief that a target is present (which are assumed to be 0.5 in all our model variants), or by assuming suboptimal or biased decision-making, but by formalizing the idea that decisions about presence are made once a target is perceived, whereas decisions about absence are made once the subject believes that the target would be perceived, if present. In doing so, the model distinguishes between two classes of parameters to describe actual visibility and beliefs about visibility, with the first set contributing mostly to decisions in the presence of a target, and the second having a stronger influence on decisions when a target is absent. With these parameters, the model further accounts for reliable heterogeneity in the effect of partial occlusion on individual participants' decisions as revealing differences in metacognitive beliefs about the manipulation, or in the tendency to incorporate these beliefs into the perceptual decision-making process.

Importantly, our model cannot decide between these two equally valid interpretations. Indeed, knowing that occlusion affects stimulus visibility is a precondition for rationally incorporating this knowledge into the decision-making process. While inaccurate meta-perceptual knowledge about occlusion may seem like an unlikely account, an increase in the effect of occlusion on the false alarm rate between Experiments 2 and 3 indicates that having direct access to the effects of occlusion on stimulus visibility does facilitate the reliance on expected visibility in perception. Importantly, however, the fact that target-absent decision times in Experiment 3 were subject to the same population variability, with some participants making reliably faster decisions about absence when more of the display was occluded, suggests that incomplete meta-perceptual knowledge cannot be the

full explanation. Relatedly, while our full model accounted for the group null effect of occlusion on target-absent decision times, it predicted a similar group-level averaging-out of the effect of occlusion on the false alarm rates. This suggests that factors other than the ones included in our model may underlie some population variability in decision times. One such factor is variability in the shape, rather than the steepness, of the temporal discounting curve. It is conceivable that variability in the temporal planning horizon, or in the degree to which value is discounted hyperbolically rather than exponentially, may underlie some variability in the effect of occlusion on target-absent RTs. For example, an impatient, hyperbolic discounter may be less likely to wait to accumulate more evidence when the expected rate of evidence accumulation is low, exactly because they know that a target is less likely to be perceived, if present.

Our findings fit with inferential, "inverse optics" accounts of perception, according to which vision is the inversion of an internal generative model of how world states translate to perceptual states, in light of noisy sensory data (Friston, 2010; Gershman et al., 2012; Smith, 2001; von Helmholtz, 1866). Given our participants' overall successful incorporation of beliefs about occlusion into perceptual decisions, previous reports of a failure to adjust a detection criterion as a function of expected visibility are more likely to reflect limited meta-perceptual knowledge (e.g., of the lower vision acuity in the visual periphery, Solovey et al., 2015), or a limited ability to use recently acquired knowledge in a flexible manner (Gorea & Sagi, 2000), rather than a blanket inability to incorporate beliefs about the perceptual evidence-generating model into perceptual decisions.

Our new model raises some questions for the ways in which perceptual detection decisions are typically modeled and conceptualized. Drift diffusion models, for example, often conceive of presence and absence evidence accumulation as a symmetric process, where evidence is similarly accumulated for the two options (Gold & Shadlen, 2001; Ratcliff et al., 2016; Ratcliff & McKoon, 2008). Moreover, the parameters of such models—the drift rate, starting point bias, and the boundary separation—do not explicitly correspond to the agent's beliefs and preferences, but to the nuts and bolts of the decision-making process (which can be linked to beliefs and desires, assuming optimal decision-making, Moran, 2015). Our ideal observer model provides such a description in terms of the subject's beliefs and goals, allowing for a separation between the perceptual likelihood function and the agents' beliefs about this function. Our findings suggest that these beliefs must be incorporated into models of perceptual decisions, and that they affect not only participants' decision threshold but also the interpretation of incoming evidence. We provide a model and accompanying code which makes explicit the separation between perceptual input and its interpretation, with the hope that it can prove useful in modeling of perceptual tasks more broadly.

Casting perceptual decisions as the result of optimal decision-making in a partially observed environment is a promising avenue for future research. This approach opens up opportunities to precisely quantify aspects of perceptual decisions which were beyond our focus here. For example, while in our model the probability of sensor activation was a function of target presence and occlusion level alone, future work may seek to identify a mapping between sensor activation probabilities and fluctuations in stimulus contrast or noise levels. Furthermore, manipulations of prior beliefs about the probability of target presence (e.g., using cuing or base-rate paradigms) can be modeled directly as the agent's beliefs about the external environment.

Decision time profiles may be better accounted for by modeling temporal discounting using a hyperbolic or a finite-horizon function—an approach which proved useful in the modeling of planning outside perception (Ainslie, 2017). Finally, the number of sensors and their sensitivity profiles can be allowed to vary, capturing settings in which evidence for one perceptual category is less informative than for the other, rather than fully missing.

We see this work as establishing that many—but possibly not all—individuals use beliefs about expected visibility to inform their perceptual decisions, and that traces of these beliefs can be identified in the way detection decisions are made, especially in the absence of a target. For our purpose here, we used the most obvious visibility manipulation that we could think of: partial stimulus occlusion. A promising direction for future research would be to use target-absent trials to infer what people know and believe about their own perception. For example, it may turn out that some manipulations affect visibility without affecting expected visibility, or vice versa: affecting expected visibility with no real effect on visibility. This way, inferences in the absence of a target can provide a window into implicit metacognitive knowledge about perception (Mazor, 2021).

Finally, we find robust individual differences in the way individuals made inferences based on the absence of perceptual evidence, with a subset of participants who systematically speed up, and make more “target-absent” judgments, with higher levels of occlusion. One possibility is that these participants committed less effort, adopting the rule-based IGNORE policy instead of the more effortful INCORP policy. While a low-effort account may explain the behavior of some participants, a null correlation between task accuracy and IGNORE-like RT patterns makes a pure effort-based explanation less likely. An alternative is that instead of reporting their best guess regarding the objective presence of a letter behind the occluders, some participants reported the subjective presence of a percept of a letter as experienced from their perspective. Indeed, if the task is to report whether a target is seen rather than whether a target is present, high occlusion should make target-absent decisions very easy. In our instructions, we tried to encourage an “external reality” reading, asking participants to “ignore the black lines, focus on the noise that is behind it, and determine whether the letter appeared in it or not.” Furthermore, trial-wise feedback about objective accuracy throughout the whole experiment was included to indicate to participants that it is the state of the world, not their subjective percept, that they are asked about. Still, some participants may have reported their subjective experience rather than their best guess of the objective reality. Interestingly, according to this interpretation, only participants who attempted to answer a question about the external world needed to “invert optics” and take their beliefs about their own perception into account. The IGNORE policy, on the other hand, is perfectly consistent with reporting what one subjectively perceives.

Two additional accounts of this individual variability identify its origins in broader principles of cognition and perception: counterfactual reasoning and probabilistic perceptual inference. First, being able to reach a conclusion early based on the absence of expected evidence is a form of counterfactual reasoning: “I would have seen the target by now if it was present.” A tendency to consider counterfactuals in perceptual inferences may be a specific instance of a domain-general individual tendency to consider counterfactuals more broadly, for example, in making inferences based on vignettes (Byrne & Tasso, 1999), or from the absence of evidence (Hsu et al., 2017). And second, the incorporation of beliefs about the perceptual likelihood

function may covary with susceptibility to other effects of beliefs on perception, for example, in cue–stimulus conditioning (Kok et al., 2013; Powers et al., 2017; Press et al., 2020), suggesting that beliefs about the likelihood and content of sensations are mediated by similar mechanisms. In both cases, robust individual variability raises the question of whether a failure to incorporate meta-perceptual beliefs is always maladaptive, or can sometimes be rational, for example, in settings where the mapping from world states to sensory input is unpredictable or unstable. We are eager to find out the answers to this and similar questions as future research will elucidate the theoretical significance of these individual differences.

## Conclusion

Overall, analysis of decision criteria, RTs and decision confidence, followed by a more focused examination of individual differences, indicates that people generally take into account beliefs about visibility when making perceptual detection judgments and when rating their subjective confidence in such decisions. Furthermore, the incorporation of beliefs about visibility into perceptual decisions is subject to substantial variability, independent of variability in physical visibility itself. This variability is especially evident in decisions when a target is absent. Our novel model of perceptual detection fits these data and provides a useful tool that importantly extends current models of perceptual decisions to incorporate beliefs not only about the world, but, crucially, about perception itself.

## Extended Method

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. Experiments 1, 2, and 3 correspond to Experiments 3, 4, and 6 of a project looking at the effects of different manipulations on inference about absence. Experiments 1, 2, and 5 used a context manipulation and will be reported separately.

Experiments 1, 2, and 3 were preregistered prior to data collection (Experiment 1: <https://osf.io/e6x82>, Experiment 2: <https://osf.io/5yr9e>, Experiment 3: <https://osf.io/mfd2w>). The long versions of Experiments 2 and 3 were not preregistered. To ensure preregistration time-locking (in other words, that preregistration preceded data collection), we employed randomization-based preregistration. We used the SHA256 cryptographic hash function to translate our preregistered protocol folders (Experiment 1: <https://github.com/matanmazor/counterfactualVisibility/blob/main/experiments/Exp1pixels/version2/protocolFolder.zip>; Experiment 2: <https://github.com/seLf-model/counterfactualVisibility/blob/main/experiments/Exp2rows/protocolFolder.zip>; Experiment 3: <https://github.com/matanmazor/counterfactualVisibility/blob/main/experiments/Exp3reference/protocolFolder.zip>) to strings of 256 bits (protocol sums; Experiment 1: e420455976659d9a46582ea0f7a64ba9e33810d90786c5157e2a188e8dcdd7c0; Experiment 2: bf72004d226b7a89a2085b0d6238a8d9b9c638513127a47fd44c6a7d00112b2f; Experiment 3: 2be4e2548db0a221a06c936fbba47cecd28894e0400477ac4f580222b77a4a44). These bits were then combined with the unique identifiers of single subjects, and the resulting string was used as seed for initializing the Mersenne Twister pseudorandom number generator prior to determining the order and timing of experimental events. This way, experimental randomization was causally dependent on, and,

therefore, could not have been determined prior to, the specific contents of our preregistration document (Mazor et al., 2019).

## Participants

The research complied with all relevant ethical regulations and was approved by the Research Ethics Committee of Birkbeck, University of London (study ID number 1812000). In all experiments, participants were recruited via Prolific and gave informed consent prior to their participation. To be eligible to take part, their Prolific approval rate had to be 95% or higher, their reported first language English, and their age between 18 and 60. Our preregistered plan was to collect data until we reach 250, 210, and 250 participants for Experiments 1, 2, and 3, respectively. Due to an error in the preprocessing script, we ultimately collected data from 234 included participants (after applying our preregistered exclusion criteria) in Experiment 2. We opted to keep the additional participants, noting that their inclusion does not change the pattern of the results. The experiments took ~12 min to complete, and participants were paid according to an hourly wage of £7.50.

For the long versions of Experiments 2 and 3, we contacted all participants who had accuracy of 70% or higher, and who did not require more than one iteration over the instructions before passing the comprehension check. The first 10 participants from each study to accept our invitation were invited to take part in five 20-min experiments, which they could complete in their own free time.

## Procedure

Participants detected the presence or absence of a target letter (S or A, in different blocks) in a patch of dynamic grayscale noise presented at 15 frames per second. In each frame, noise was generated by randomly sampling grayscale values from a target image  $I$ . Specifically, for each pixel  $S_{ij}$ , we displayed the grayscale value for the corresponding pixel in the original, noise-free, image  $I_{ij}$  with some probability  $p$ , and the grayscale value of a randomly chosen pixel  $I_{ij'}$  (sampled with replacement) with probability  $1 - p$ . On target-absent trials,  $p$  was set to 0, such that grayscale values of all pixels were randomly shuffled, with replacement. On target-present trials, the probability  $p$  was set to a positive number between 0 and 1. In Experiments 1 and 2,  $p$  was calibrated online to achieve performance levels of around 80%, following a one-up-three-down procedure, starting at  $v = 0.35$  and following a multiplicative set size of 0.9, which moved closer to 1 following each change direction in the calibration process. In Experiment 3,  $p$  was set to .3 throughout the entire experiment. Responses were delivered using the F and G keyboard keys (counterbalancing response mapping across subjects).

After reading the instructions, participants completed four practice trials. In case their accuracy in these four practice trials fell below 3/4, they were reminded of task instructions and given additional practice trials, until reaching the desired accuracy level. Otherwise, they continued to the main part of the experiment. Here, their task was exactly the same, but the noise patch was partly occluded. In Experiment 1, occluders were randomly positioned static black pixels, which covered 5% or 15% of the stimulus on different trials. In Experiments 2 and 3, occluders were randomly positioned rows of black pixels (two or six rows, on different trials)

which extended beyond the stimulus. In order to make clear that the occluders are not part of the main stimulus, occluder rows in Experiments 2 and 3 preceded the main stimulus by 500 ms. Finally, in Experiment 3, two similar “reference” stimuli were presented on both sides of the central stimulus. In these reference stimuli, the target letter was always presented with  $p = .3$  regardless of the presence of a letter in the central stimulus. Participants were explained that they should respond based on the central stimulus only and continued to the main part of the experiment only once they had passed a comprehension check.

The main part of the experiment comprised four blocks of 16 trials. For approximately half of the participants, in Blocks 1 and 2 the target letter was S and in Blocks 3 and 4 it was A. The order of letters was reversed for the other half. In Blocks 3 and 4 of Experiment 2, participants used their mouse to rate their confidence on a vertical analog scale immediately after deciding whether the letter was present or absent. To move on to the third block, participants had to respond correctly on at least three of four trials, and to correctly answer a multiple-option comprehension question about the use of the confidence scale.

Finally, at the end of Experiment 2, participants were asked to report whether occlusion affected how difficult it was to detect the letter. 197 participants reported that occluding more of the stimulus made detecting the target letter harder, five reported it made detecting the target letter easier, and the remaining 32 reported it had no effect on difficulty.

## Statistical Analysis

All statistical tests were tested using the standard 0.05 significance threshold. Bayes factors, when reported, assume a noninformative Cauchy prior (scale factor = 0.707) over effect sizes, equivalent to a belief that an effect, when present, is similarly likely to be greater or smaller than 0.707 SDs.

## Data Exclusion

We followed our preregistered exclusion criteria. Participants were excluded if their accuracy fell below 50%, and for having extremely fast or slow RTs in more than 25% of the trials. Too fast RTs were defined as below 100 ms in all experiments. Too slow RTs were defined as above 5 s in Experiments 1 and 2, and above 7 s in Experiment 3 and in the long versions of Experiments 2 and 3.

Trials with too slow or too fast response times according to the above criteria were excluded from the response time analysis.

## Reverse Correlation Analysis

For each individual frame, we computed the correlation between pixel luminance values and the corresponding values in the target image. Occluded pixels were omitted from both stimulus and target image. This resulted in an uncorrected stimulus–target similarity metric, which was naturally higher for frames from target-present trials, and for frames with higher levels of stimulus visibility ( $p$ , see the Procedure subsection). We therefore mean-centered these correlation values as a function of stimulus presence, occlusion and  $p$ , to obtain an unbiased measure of stimulus–target similarity  $r$ .

For each individual subject, we averaged these frame-wise correlations across frames 1–5, roughly corresponding to the first 300 ms of the trial. We then calculated, for each participant, the Spearman correlation between these average values and trial-wise RTs. This was done separately for target-present and target-absent trials. These subject-level correlations were then subjected to a group-level *t* test.

In Experiment 2, a similar procedure was used to reveal correlations between stimulus-target similarity and subjective confidence. For each participant, we calculated the Spearman correlation between trial-wise values of *r*, averaged across frames 1–5, and trial-wise confidence ratings, separately for target-present and target-absent trials. We then subjected the resulting correlations to a group-level *t* test.

## Model Simulations

### Model Specification

A POMDP is a 7-tuple:  $\langle S, A, T, \Omega, O, r, \gamma \rangle$ . The state space *S* comprises two states describing target presence or absence and two additional states for trial endings: correct and incorrect. The action space *A* has three possible actions: “wait,” “decide present,” and “decide absent.” The transition function  $T: (S, A) \rightarrow S$  specifies the effect of actions on state transitions. “Wait” maps states to themselves, and deciding maps states to the terminal “correct” or “incorrect” states depending on the accuracy of the decision, which have no associated actions with them.  $\Omega$  is the set of possible observations. We assume these are  $[0, 1]$ , that is, perceptual evidence has a binary form.  $O: S \rightarrow P(\Omega)$  is a probabilistic function from states to observations, which we describe in more detail below.  $r: S \rightarrow R$  maps states to reward values. We set the values of all states to 0, except “correct” which is associated with a value of 1. Finally, the temporal discount factor  $\gamma$  affects the subjective value of anticipated rewards. We set  $\gamma: .99$  meaning that a reward obtained in the next time point is worth .99 of its worth if obtained now.

The observation function *O* is a Bernoulli function, such that the probability of observing 1 equals the bias parameter  $\theta$  which depends on target presence. Specifically, we set

$$\theta: = \begin{cases} 0.05 & \text{absent} \\ 0.2 & \text{present} \end{cases}. \quad (1)$$

Importantly, for any choice of  $\theta$  such that  $0 < \theta_{\text{absent}} < \theta_{\text{present}} < 0.5$ , positive evidence (i.e., sampling a 1) is more informative than negative evidence (i.e., sampling a 0). For example, for the values we use here, after sampling a 0 an agent should update their subjective belief that a target is present only by a small amount, from 0.5 to 0.46. In contrast, after sampling a single 1, belief update is much steeper: from 0.5 to 0.8.

Agents need to infer target presence from noisy observations. Their belief state can therefore be described as the LLR between target presence and absence, which they update following each sample.

$$\text{LLR}_t = \sum_{i=1}^t \log \frac{p(o_i | \bar{\theta}_{\text{presence}})}{p(o_i | \bar{\theta}_{\text{absence}})}, \quad (2)$$

where

$$p(o_i | \bar{\theta}) = \begin{cases} \bar{\theta} & \text{if } o_i = 1 \\ 1 - \bar{\theta} & \text{if } o_i = 0 \end{cases}, \quad (3)$$

with  $\bar{\theta}$  being the assumed value of  $\theta$  in the agent’s internal model of their perception (in all our simulations,  $\bar{\theta} = \theta$ ). The probability that a target is present given the evidence so far is then:

$$p(\text{present} | O_t) = \frac{e^{\text{LLR}_t}}{1 + e^{\text{LLR}_t}}. \quad (4)$$

With  $O_t$  being the entire stream of evidence until time point *t*. And, assuming that, at the time of committing to a decision, the agent decides “present” if and only if  $p(\text{present}|O) > .5$ , the probability of being correct at that time point is:

$$p(\text{correct} | \text{DECIDE}, O_t) = \max(p(\text{present}|O_t), 1 - p(\text{present}|O_t)), \quad (5)$$

when following the optimal policy, the expected value at time point *t* equals the maximum of (a) the probability of being correct if decision is taken now, and (b) the expected value of waiting and collecting additional evidence, discounted by the temporal discount factor  $\gamma$ :

$$E(V|O_t) = \max(p(\text{correct}|O_t), p(1|O_t)\gamma E(V|[O_t, 1]) + p(0|O_t)\gamma E(V|[O_t, 0])), \quad (6)$$

where  $E(V|[O_t, 0])$  is the expected value at time point *t* + 1, assuming the next sample is 0, and  $p(1|O_t) = p(\text{present}|O_t)\bar{\theta}_{\text{present}} + p(\text{absent}|O_t)\bar{\theta}_{\text{absent}}$  is the probability that the next sample will be 1, marginalized over target presence and absence (similar for 0). The optimal action at time *t* is determined by the maximizing term (deciding now or waiting).

Finally, confidence ratings are modeled as the estimated probability of being correct when committing to a decision.

### Occlusion Effects

We simulate stimulus occlusion as a scaling of the probability of obtaining positive evidence by a parameter  $\alpha \in [0, 1]$ . Similar to  $\theta_{\text{present}}$  and  $\theta_{\text{absent}}$ ,  $\alpha$  is paralleled by a metacognitive variable,  $\bar{\alpha}$ , which corresponds to participants’ beliefs about the effects of occlusion on stimulus visibility. This way of defining occlusion has three notable characteristics. First, the relative effect of occlusion on the probability of sampling a 1 ( $\alpha$ ) is much more pronounced than its positive effect on the probability of sampling a 0 ( $\frac{1-\alpha\theta}{1-\theta}$ ). For example, for the case of  $\theta = 0.1$  and  $\alpha = 0.7$ , occlusion reduces the probability of sampling a 1 by a factor of 1.43 but increases the probability of sampling a 0 by a factor of 1.03 only.

Second, the informativeness of obtaining positive evidence, quantified as the LLR between target presence and absence following a 1, is unaffected by beliefs about the effects of occlusion on visibility,  $\bar{\alpha}$ :

$$\text{LLR}_{[1]} = \log \frac{p(1|\text{present})}{p(1|\text{absent})} = \log \frac{\bar{\alpha}\bar{\theta}_{\text{present}}}{\bar{\alpha}\bar{\theta}_{\text{absent}}} = \log \frac{\bar{\theta}_{\text{present}}}{\bar{\theta}_{\text{absent}}}. \quad (7)$$

And third, the informativeness of obtaining negative evidence, quantified as the LLR between target presence and absence following a 0, approaches 0 with lower values of  $\bar{\alpha}$ , as if the model considers the probability that evidence would have been obtained if a target was present:

$$\begin{aligned} |\text{LLR}_{[0]}| &= \left| \log \frac{p(0|\text{present})}{p(0|\text{absent})} \right| = \\ &\left| \log \frac{1 - \bar{\alpha}\bar{\theta}_{\text{present}}}{1 - \bar{\alpha}\bar{\theta}_{\text{absent}}} \right| < \left| \log \frac{1 - \bar{\theta}_{\text{present}}}{1 - \bar{\theta}_{\text{absent}}} \right|. \end{aligned} \quad (8)$$

Together, we get a double dissociation. Occlusion affects the probability of obtaining positive evidence, but beliefs about occlusion have no effect on the interpretation of such evidence once obtained. On the other hand, occlusion has little effect on the probability of obtaining negative evidence, but beliefs about the effects of occlusion affect the interpretation of such evidence once obtained. As a result, timing and confidence in decisions about absence depend much more on beliefs about the effect of occlusion than on the true effect of occlusion on visibility.

In the simulations we had two occlusion levels; one where  $\alpha = 0.8$  (easy condition) and one where  $\alpha = 0.6$  (hard condition). We present the results of two simulated agents:  $V_{\text{INCORP}}$  is an ideal observer who uses information about the expected effect of occlusion on visibility to interpret data and make decisions ( $\bar{\alpha} = \alpha$ ), and  $V_{\text{IGNORE}}$  is an observer who interprets perceptual evidence similarly in both levels of occlusion ( $\bar{\alpha} = .7$  for both hard and easy conditions). To find the optimal policy (the one that maximizes the Bellman equation), we used backward induction with a horizon of 100 time points (Callaway et al., 2024; Puterman, 2014). We then simulated 4,000 trials to obtain predictions for a rational decision-maker, for each agent separately.

## Model Fitting

Model parameters were fitted to the behavior of individual participants. Ten model parameters were included:

1.  $\theta_{\text{absent}} = p(1|\text{absent})$
2.  $\theta_{\Delta} = p(1|\text{present}) - p(1|\text{absent})$
3.  $\bar{\theta}_{\text{absent}}$
4.  $\bar{\theta}_{\Delta}$

Parameters 1–4 were allowed to vary between 0.00005 and 0.27.

5.  $\gamma$ : the temporal discounting parameter. Allowed to vary between 0.989 and 0.99995.
6. Minimal nondecision time. Allowed to vary between 0.2 and 1 s.
7. Maximal nondecision time minus minimal nondecision time. Allowed to vary between 0.1 and 1 s.
8.  $\alpha$ : the effect of occlusion on visibility. For model-fitting purposes,  $p(1)$  in the easy condition was set to  $\theta/\alpha$  and in the hard condition  $\theta\alpha$ .

9.  $\bar{\alpha}$ : the believed effect of occlusion on visibility. For model-fitting purposes, here  $\bar{p}(1)$  in the easy condition was  $\bar{\alpha}/\theta$  and in the hard condition  $\bar{\theta}/\bar{\alpha}$ .

Parameters 8 and 9 were allowed to vary between 0.67 and 1. To account for noise in the decision-making process, action selection followed a softmax distribution:

$$p(a) = \frac{\exp\left(\frac{v_a}{T}\right)}{\sum_a \exp\left(\frac{v_a}{T}\right)}, \quad (9)$$

where  $v_a$  is the value associated with taking action  $a$ , and  $T$  is the softmax temperature:

10.  $T$ . Allowed to vary between 0.0015 and 1. Notably, the model fits of most participants converged to the lowest possible value for  $T$ , indicating very little decision noise.

To aid with model fitting, parameters 1–5 were fitted in logit space and then transformed via a sigmoid function, and parameters 8–10 were fitted in log space and then exponentiated.

Model fitting was carried out in Matlab (version R2023a, Optimization Toolbox). We used a combination of simulated annealing (Matlab's simannealbnd) and the nonlinear programming solver fmincon. For Experiments 1–3, we ran 12 independent optimizations per participant, starting at random points in the parameter space, and used the parameters that produced the best fits in terms of log likelihood. For the long Experiments 2a and 2b, we ran 48 independent optimizations per participant. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing facility in carrying out this work (<http://doi.org/10.5281/zenodo.22558>).

## Model Variants

In addition to fitting the model as described, we fitted three additional model variants:

### An IGNORE Variant

In this version,  $\bar{\alpha}$  was set to 1.

### An INCORP Variant

In this version,  $\bar{\alpha}$  was set to be equal to  $\alpha$ .

### A Symmetric Variant

In this version, the agent had access to two sensors,  $P$  and  $A$ , that are symmetrically tuned to the presence or absence of a target. As a result, momentary individual observations were now pairs of values  $[o^P, o^A]$ : [1,1], [0,0], [1,0] or [0, 1]. The probability of the presence sensor to activate when a target was present,  $\theta_{\text{present}}^P$  was identical to the probability of the absence sensor to activate when no target is present  $\theta_{\text{absent}}^A$ . Similarly,  $\theta_{\text{absent}}^P = \theta_{\text{present}}^A$ . The observers' beliefs were symmetric in the same way:  $\theta_{\text{present}}^P = \bar{\theta}_{\text{absent}}^A$  and  $\bar{\theta}_{\text{present}}^P = \bar{\theta}_{\text{absent}}^A$ . This is an implementation of the neuron–antineuron architecture, described in Gold and Shadlen's (2001) as an alternative to likelihood-based inference. More practically, the first

four parameters in the original model were replaced with the following parameters:

1.  $\theta_{\text{unpreferred}} = p(o^P = 1|\text{absent}) = p(o^A = 1|\text{present})$
2.  $\theta_\Delta = p(1|\text{preferred}) - p(1|\text{unpreferred})$
3.  $\bar{\theta}_{\text{unpreferred}}$
4.  $\theta_\Delta$

where  $\theta_{\text{preferred}} = p(o^A = 1|\text{absent}) = p(o^P = 1|\text{present})$ . The agent updates their beliefs and determines the optimal policy given these observations, and using the same procedure used in the asymmetric model.

## References

- Ainslie, G. (2017). De gustibus disputare: Hyperbolic delay discounting integrates five approaches to impulsive choice. *Journal of Economic Methodology*, 24(2), 166–189. <https://doi.org/10.1080/1350178X.2017.1309373>
- Altman, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *The BMJ*, 311(7003), Article 485. <https://doi.org/10.1136/bmj.311.7003.485>
- Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press. <https://doi.org/10.4324/9780203771730>
- Beltrán, D., Liu, B., & de Vega, M. (2021). Inhibitory mechanisms in the processing of negations: A neural reuse hypothesis. *Journal of Psycholinguistic Research*, 50(6), 1243–1260. <https://doi.org/10.1007/s10936-021-09796-x>
- Byrne, R. M. J., & Tasso, A. (1999). Deductive reasoning with factual, possible, and counterfactual conditionals. *Memory & Cognition*, 27(4), 726–740. <https://doi.org/10.3758/BF03211565>
- Calder-Travis, J., Charles, L., Bogacz, R., & Yeung, N. (2024). Bayesian confidence in optimal decisions. *Psychological Review*, 131(5), 1114–1160. <https://doi.org/10.1037/rev0000472>
- Callaway, F., Griffiths, T. L., Norman, K. A., & Zhang, Q. (2024). Optimal metacognitive control of memory recall. *Psychological Review*, 131(3), 781–811. <https://doi.org/10.1037/rev0000441>
- Chun, M. M., & Wolfe, J. M. (1996). Just say no: How are visual searches terminated when there is no target present? *Cognitive Psychology*, 30(1), 39–78. <https://doi.org/10.1006/cogp.1996.0002>
- Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., & Powers, A. R., III. (2019). Hallucinations and strong priors. *Trends in Cognitive Sciences*, 23(2), 114–127. <https://doi.org/10.1016/j.tics.2018.12.001>
- Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for “top-down” effects. *Behavioral and Brain Sciences*, 39, Article e229. <https://doi.org/10.1017/S0140525X15000965>
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138. <https://doi.org/10.1038/nrn2787>
- Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, 24(1), 1–24. [https://doi.org/10.1162/NECO\\_a\\_00226](https://doi.org/10.1162/NECO_a_00226)
- Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, 5(1), 10–16. [https://doi.org/10.1016/S1364-6613\(00\)01567-9](https://doi.org/10.1016/S1364-6613(00)01567-9)
- Goldberg, S. (2011). If that were true I would have heard about it by now. In A. I. Goldman & D. Whitcomb (Eds.), *Social epistemology: Essential readings* (pp. 92–108). Oxford University Press.
- Gorea, A., & Sagi, D. (2000). Failure to handle more than one internal representation in visual detection tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(22), 12380–12384. <https://doi.org/10.1073/pnas.97.22.12380>
- Haarsma, J., Deveci, N., Corbin, N., Callaghan, M. F., & Kok, P. (2023). Expectation cues and false percepts generate stimulus-specific activity in distinct layers of the early visual cortex. *The Journal of Neuroscience*, 43(47), 7946–7957. <https://doi.org/10.1523/JNEUROSCI.0998-23.2023>
- Hellmann, S., Zehetleitner, M., & Rausch, M. (2023). Simultaneous modeling of choice, confidence, and response time in visual perception. *Psychological Review*, 130(6), 1521–1543. <https://doi.org/10.1037/re0000411>
- Herce Castañón, S., Moran, R., Ding, J., Egner, T., Bang, D., & Summerfield, C. (2019). Human noise blindness drives suboptimal cognitive inference. *Nature Communications*, 10(1), Article 1719. <https://doi.org/10.1038/s41467-019-09330-7>
- Hsu, A. S., Horng, A., Griffiths, T. L., & Chater, N. (2017). When absence of evidence is evidence of absence: Rational inferences from absent data. *Cognitive Science*, 41(Suppl. 54), 1155–1167. <https://doi.org/10.1111/cogs.12356>
- Kellij, S., Fahrenfort, J., Lau, H., Peters, M. A. K., & Odegaard, B. (2021). An investigation of how relative precision of target encoding influences metacognitive performance. *Attention, Perception, & Psychophysics*, 83(1), 512–524. <https://doi.org/10.3758/s13414-020-02190-0>
- Ko, Y., & Lau, H. (2012). A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, 367(1594), 1401–1411. <https://doi.org/10.1098/rstb.2011.0380>
- Kok, P., Brouwer, G. J., van Gerven, M. A. J., & de Lange, F. P. (2013). Prior expectations bias sensory representations in visual cortex. *The Journal of Neuroscience*, 33(41), 16275–16284. <https://doi.org/10.1523/JNEUROSCI.0742-13.2013>
- Littman, M. L. (2009). A tutorial on partially observable markov decision processes. *Journal of Mathematical Psychology*, 53(3), 119–125. <https://doi.org/10.1016/j.jmp.2009.01.005>
- Locke, J. (1690). *An essay concerning human understanding*. Thomas Bassett. <https://doi.org/10.1093/oseo/instance.00018020>
- Ma, W. J. (2012). Organizing probabilistic models of perception. *Trends in Cognitive Sciences*, 16(10), 511–518. <https://doi.org/10.1016/j.tics.2012.08.010>
- Ma, W. J. (2019). Bayesian decision models: A primer. *Neuron*, 104(1), 164–175. <https://doi.org/10.1016/j.neuron.2019.09.037>
- Maloney, L. T., & Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience*, 26(1), 147–155. <https://doi.org/10.1017/S0952523808080805>
- Mamassian, P., Landy, M., & Maloney, L. T. (2002). Bayesian modelling of visual perception. In R. P. N. Rao, B. A. Olshausen, & M. S. Lewicki (Eds.), *Probabilistic models of the brain: Perception and neural function* (pp. 13–36). The MIT Press. <https://doi.org/10.7551/mitpress/5583.003.0005>
- Mazor, M. (2021). *Inference about absence as a window into the mental self-model*. <https://doi.org/10.31234/osf.io/zgf6s>
- Mazor, M., Friston, K. J., & Fleming, S. M. (2020). Distinct neural contributions to metacognition for detecting, but not discriminating visual stimuli. *eLife*, 9, Article e53900. <https://doi.org/10.7554/eLife.53900>
- Mazor, M., Maimon-Mor, R. O., Charles, L., & Fleming, S. M. (2023). Paradoxical evidence weighting in confidence judgments for detection and discrimination. *Attention, Perception, & Psychophysics*, 85(7), 2356–2385. <https://doi.org/10.3758/s13414-023-02710-8>
- Mazor, M., Mazor, N., & Mukamel, R. (2019). A novel tool for time-locking study plans to results. *European Journal of Neuroscience*, 49(9), 1149–1156. <https://doi.org/10.1111/ejn.14278>
- Mazor, M., Moran, R., & Fleming, S. M. (2021). Erratum to: Stage 1 registered report: Metacognitive asymmetries in visual perception and Stage 2 registered report: Metacognitive asymmetries in visual perception.

- Neuroscience of Consciousness*, 2021(1), Article niab046. <https://doi.org/10.1093/nc/niab046>
- Meuwese, J. D. I., van Loon, A. M., Lamme, V. A. F., & Fahrenfort, J. J. (2014). The subjective experience of object recognition: Comparing metacognition for object detection and object categorization. *Attention, Perception, & Psychophysics*, 76(4), 1057–1068. <https://doi.org/10.3758/s13414-014-0643-1>
- Moran, R. (2015). Optimal decision making in heterogeneous and biased environments. *Psychonomic Bulletin & Review*, 22(1), 38–53. <https://doi.org/10.3758/s13423-014-0669-3>
- Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, 78, 99–147. <https://doi.org/10.1016/j.cogpsych.2015.01.002>
- Oaksford, M., & Hahn, U. (2004). A bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale*, 58(2), 75–85. <https://doi.org/10.1037/h0085798>
- Olawole-Scott, H., & Yon, D. (2023). Expectations about precision bias metacognition and awareness. *Journal of Experimental Psychology: General*, 152(8), 2177–2189. <https://doi.org/10.1037/xge0001371>
- Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. *Psychonomic Bulletin & Review*, 10(1), 177–183. <https://doi.org/10.3758/BF03196482>
- Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. *Science*, 357(6351), 596–600. <https://doi.org/10.1126/science.aan3458>
- Press, C., Kok, P., & Yon, D. (2020). The perceptual prediction paradox. *Trends in Cognitive Sciences*, 24(1), 13–24. <https://doi.org/10.1016/j.tics.2019.11.003>
- Puterman, M. L. (2014). *Markov decision processes: Discrete stochastic dynamic programming*. Wiley.
- Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, 22(3), 341–365. <https://doi.org/10.1017/S0140525X99002022>
- Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20(4), 873–922. <https://doi.org/10.1162/neco.2008.12-06-420>
- Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current issues and history. *Trends in Cognitive Sciences*, 20(4), 260–281. <https://doi.org/10.1016/j.tics.2016.01.007>
- Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception, & Psychophysics*, 80(1), 134–154. <https://doi.org/10.3758/s13414-017-1431-5>
- Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, 5(2), 97–118. <https://doi.org/10.1080/17588928.2013.877880>
- Smith, A. M. (2001). *Alhacen's theory of visual perception: A critical edition, with English translation and commentary, of the first three books of Alhacen's De Aspectibus, the Medieval Latin Version of Ibn Al-Haytham's Kitab Al-Manazir*. American Philosophical Society.
- Solovey, G., Graney, G. G., & Lau, H. (2015). A decisional account of subjective inflation of visual perception at the periphery. *Attention, Perception, & Psychophysics*, 77(1), 258–271. <https://doi.org/10.3758/s13414-014-0769-1>
- Stelzer, J., Chen, Y., & Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage*, 65, 69–82. <https://doi.org/10.1016/j.neuroimage.2012.09.063>
- Stuke, H., Weilnhammer, V. A., Sterzer, P., & Schmack, K. (2019). Delusion proneness is linked to a reduced usage of prior beliefs in perceptual decisions. *Schizophrenia Bulletin*, 45(1), 80–86. <https://doi.org/10.1093/schbul/sbx189>
- Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, 13(9), 403–409. <https://doi.org/10.1016/j.tics.2009.06.003>
- Tajima, S., Drugowitsch, J., & Pouget, A. (2016). Optimal policy for value-based decision-making. *Nature Communications*, 7(1), Article 12400. <https://doi.org/10.1038/ncomms12400>
- Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review*, 91(1), 68–111. <https://doi.org/10.1037/0033-295X.91.1.68>
- von Helmholz, H. (1866/2009). Concerning the perceptions in general. In James P. C. Southallno (Ed.), *Treatise on physiological optics*. Dover.
- Walton, D. (1992). Nonfallacious arguments from ignorance. *American Philosophical Quarterly*, 29(4), 381–387. <https://www.jstor.org/stable/20014433>
- Walton, D. (2010). *Arguments from ignorance*. Penn State Press.
- Weilnhammer, V. A., Stuke, H., Sterzer, P., & Schmack, K. (2018). The neural correlates of hierarchical predictions for perceptual decisions. *The Journal of Neuroscience*, 38(21), 5008–5021. <https://doi.org/10.1523/JNEUROSCI.2901-17.2018>
- Yaron, I., Faivre, N., Mudrik, L., & Mazor, M. (2023). *Individual differences do not mask effects of unconscious processing*. <https://doi.org/10.31234/osf.io/ebg8w>
- Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, 367(1594), 1310–1321. <https://doi.org/10.1098/rstb.2011.0416>
- Yon, D., & Frith, C. D. (2021). Precision and the Bayesian brain. *Current Biology*, 31(17), R1026–R1032. <https://doi.org/10.1016/j.cub.2021.07.044>
- Yon, D., Thomas, E. R., Gilbert, S. J., de Lange, F. P., Kok, P., & Press, C. (2023). Stubborn predictions in primary visual cortex. *Journal of Cognitive Neuroscience*, 35(7), 1133–1143. [https://doi.org/10.1162/jocn\\_a\\_01997](https://doi.org/10.1162/jocn_a_01997)
- Yon, D., Zainzinger, V., de Lange, F. P., Eimer, M., & Press, C. (2021). Action biases perceptual decisions toward expected outcomes. *Journal of Experimental Psychology: General*, 150(6), 1225–1236. <https://doi.org/10.1037/xge0000826>
- Zang, A., de Vega, M., Fu, Y., Wang, H., & Beltrán, D. (2022). Language switching may facilitate the processing of negative responses. *Frontiers in Psychology*, 13, Article 906154. <https://doi.org/10.3389/fpsyg.2022.906154>
- Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, 6, Article 79. <https://doi.org/10.3389/fint.2012.00079>
- Zylberberg, A., Wolpert, D. M., & Shadlen, M. N. (2018). Counterfactual reasoning underlies the learning of priors in decision making. *Neuron*, 99(5), 1083–1097.e6. <https://doi.org/10.1016/j.neuron.2018.07.035>

(Appendix follows)

## Appendix

### Preregistered Analysis

Probability correct was 0.81 ( $SD = 0.04$ ) in Experiment 1, 0.81 ( $SD = 0.03$ ) and 0.82 ( $SD = 0.10$ ) in Experiment 3.

#### **Hypothesis 1 (Presence/Absence Response Time)**

##### **Experiment 1**

A paired  $t$  test on the median individual-level response times revealed a significant difference between target-present and target-absent response times,  $M = 454.16$ , 95% CI [402.75, 505.58],  $t(250) = 17.40$ ,  $p < .001$ .

##### **Experiment 2**

A paired  $t$  test on the median individual-level response times revealed a significant difference between target-present and target-absent response times,  $M = 384.93$ , 95% CI [337.23, 432.62],  $t(233) = 15.90$ ,  $p < .001$ .

##### **Experiment 3**

A paired  $t$  test on the median individual-level response times revealed a significant difference between target-present and target-absent response times,  $M = 581.37$ , 95% CI [503.16, 659.58],  $t(248) = 14.64$ ,  $p < .001$ .

#### **Hypothesis 2 (Occlusion Effect in Presence)**

##### **Experiment 1**

A paired  $t$  test on the median individual-level response times in hit trials revealed a significant effect of occlusion on RT,  $M = -65.95$ , 95% CI [-102.95, -28.95],  $t(250) = -3.51$ ,  $p < .001$ .

##### **Experiment 2**

A paired  $t$  test on the median individual-level response times in hit trials revealed a significant effect of occlusion on RT,  $M = -134.34$ , 95% CI [-185.95, -82.73],  $t(233) = -5.13$ ,  $p < .001$ .

##### **Experiment 3**

A paired  $t$  test on the median individual-level response times in hit trials revealed a significant effect of occlusion on RT,  $M = -101.04$ , 95% CI [-139.35, -62.73],  $t(248) = -5.19$ ,  $p < .001$ .

#### **Hypothesis 3 (Occlusion Effect in Absence)**

##### **Experiment 1**

A paired  $t$  test on the median individual-level response times in correct rejection trials revealed no significant effect of occlusion on RT,  $M = 5.88$ , 95% CI [-32.81, 44.56],  $t(250) = 0.30$ ,  $p = .765$ .

##### **Experiment 2**

A paired  $t$  test on the median individual-level response times in correct rejection trials revealed no significant effect of occlusion on RT,  $M = 15.03$ , 95% CI [-22.34, 52.40],  $t(233) = 0.79$ ,  $p = .429$ .

##### **Experiment 3**

A paired  $t$  test on the median individual-level response times in correct rejection trials revealed no significant effect of occlusion on RT,  $M = 19.65$ , 95% CI [-37.33, 76.63],  $t(248) = 0.68$ ,  $p = .498$ .

#### **Hypothesis 4 (Occlusion Response Interaction)**

##### **Experiment 1**

We find a significant interaction between occlusion level and response on reaction times,  $M = -71.82$ , 95% CI [-127.16, -16.48],  $t(250) = -2.56$ ,  $p = .011$ .

##### **Experiment 2**

We find a significant interaction between occlusion level and response on reaction times,  $M = -149.37$ , 95% CI [-212.21, -86.53],  $t(233) = -4.68$ ,  $p < .001$ .

##### **Experiment 3**

We find a significant interaction between occlusion level and response on reaction times,  $M = -120.69$ , 95% CI [-189.53, -51.85],  $t(248) = -3.45$ ,  $p < .001$ .

#### **Hypothesis 5 (Sensitivity)**

##### **Experiment 1**

We find a significant drop in perceptual sensitivity ( $d'$ ) as a function of occlusion,  $M = 0.37$ , 95% CI [0.27, 0.46],  $t(250) = 7.64$ ,  $p < .001$ .

##### **Experiment 2**

We find a significant drop in perceptual sensitivity ( $d'$ ) as a function of occlusion,  $M = 0.51$ , 95% CI [0.42, 0.61],  $t(233) = 10.46$ ,  $p < .001$ .

##### **Experiment 3**

We find a significant drop in perceptual sensitivity ( $d'$ ) as a function of occlusion,  $M = 0.58$ , 95% CI [0.50, 0.67],  $t(248) = 13.64$ ,  $p < .001$ .

### Hypothesis 5 (Criterion)

#### Experiment 1

We find a significant increase in the signal detection criterion ( $c$ ) as a function of occlusion,  $M = 0.14$ , 95% CI [0.09, 0.19],  $t(250) = 5.45$ ,  $p < .001$ .

#### Experiment 2

We find a significant increase in the signal detection criterion ( $c$ ) as a function of occlusion,  $M = 0.15$ , 95% CI [0.10, 0.21],  $t(233) = 5.85$ ,  $p < .001$ .

#### Experiment 3

We find a significant increase in the signal detection criterion ( $c$ ) as a function of occlusion,  $M = 0.10$ , 95% CI [0.05, 0.15],  $t(248) = 4.02$ ,  $p < .001$ .

### Hypothesis 7 (Presence/Absence Confidence)

#### Experiment 2

A paired  $t$  test on the mean individual-level confidence ratings from correct responses only revealed a significant effect of target presence on confidence,  $M = 0.01$ , 95% CI [0.00, 0.03],  $t(233) = 2.92$ ,  $p = .004$ .

### Hypothesis 8 (Occlusion Confidence Effect in Presence)

#### Experiment 2

A paired  $t$  test on the mean individual-level confidence ratings in correct trials only revealed a significant effect of occlusion on hit confidence ratings,  $M = 0.04$ , 95% CI [0.03, 0.05],  $t(233) = 9.87$ ,  $p < .001$ .

### Hypothesis 9 (Occlusion Confidence Effect in Absence)

#### Experiment 2

A paired  $t$  test on the mean individual-level confidence ratings in correct trials only revealed a significant effect of occlusion on correct-rejection confidence ratings,  $M = 0.04$ , 95% CI [0.03, 0.05],  $t(233) = 10.54$ ,  $p < .001$ .

### Hypothesis 10 (Occlusion Response Interaction on Confidence)

#### Experiment 2

We find no significant interaction effect between occlusion and target presence on confidence,  $M = 0.00$ , 95% CI [-0.01, 0.01],  $t(233) = 0.01$ ,  $p = .992$ .

### Model Recovery

As described in the text, four models were fitted to the data from the two long experiments: a full “inverse optics” model where visibility and beliefs about visibility are independently free to vary, an “INCORP” model that assumes perfectly accurate beliefs about the

**Figure A1**  
Model Recovery Results, Experiments 2 and 3 (Long Versions)

	symmetric	1	0	0	16
Generating model					
INCORP	6	1	10	0	
IGNORE	4	12	1	0	
full	7	3	3	4	
	full	IGNORE	INCORP	symmetric	Best fitting model

Note. Confusion matrix between the four model variants. Data were generated using the parameters fitted to the 17 participants from Experiments 2 and 3 (long versions) whose fitted parameters varied between all four model variants. We then fitted the same four model variants to the generated data and recorded the best fitting model as decided with AIC. Exp. = experiment. INCORP = incorporate; AIC = Akaike Information Criterion. See the online article for the color version of this figure.

effect of occlusion on visibility, an “IGNORE” model that assumes no knowledge about the effect of occlusion on visibility, and a symmetric model that is equipped with an absence sensor in addition to a presence sensor. We used the best fitting parameters from each model to simulate 896 trials per participant, and then fitted the four models to the resulting simulated data, excluding three participants whose best fitting parameters in the full model were consistent with the IGNORE variant ( $\bar{\alpha} = 1$ ). This procedure ensured that above-chance model recovery is not driven by a choice of unrealistic model parameters.

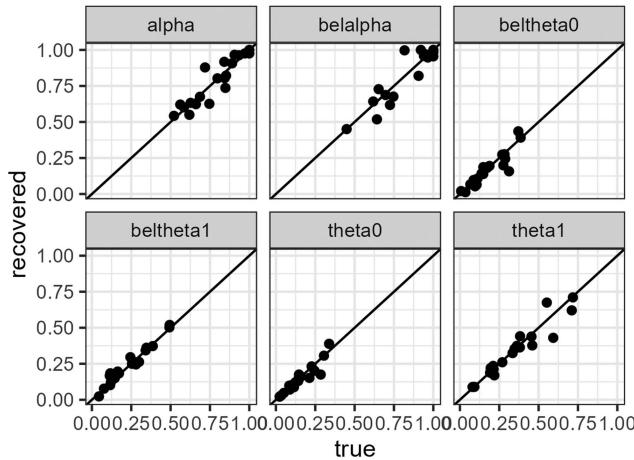
The resulting confusion matrix shows minimal confusion between the symmetric and the full model variants and between the IGNORE and INCORP variants in particular.

### Parameter Recovery

To verify that model parameters are in principle recoverable, we used the parameters fitted to participants in the long version of Experiments 2 and 3 and simulated 896 trials per participant. We then repeated the same model fitting procedure on these recovered parameters. The correlations between fitted and recovered parameters were generally very high for the parameters of interest (see Figure A2). Most importantly for our purpose, the difference between  $\alpha$  and  $\bar{\alpha}$ , indicating whether participants overestimated, underestimated, or accurately estimated the effect of occlusion on target visibility, was highly recoverable,  $r = .93$ , 95% CI [.84, .97],  $t(18) = 11.10$ ,  $p < .001$ . Furthermore, the intercept term in a linear regression model fitted to

**Figure A2**

Parameter Recovery Results, Experiments 2 and 3 (Long Versions)



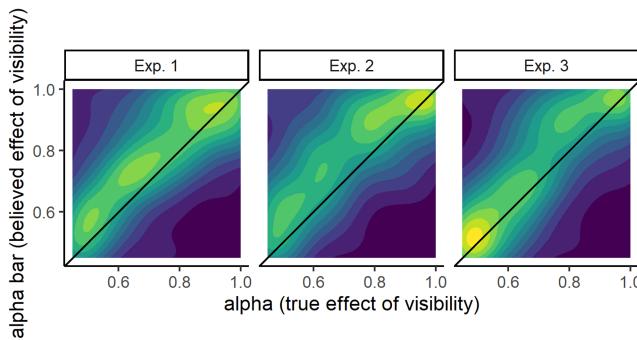
predict the recovered  $\bar{\alpha} - \alpha$  from the true  $\bar{\alpha} - \alpha$  was not statistically different from 0,  $t(18) = -0.30, p = .768$ , suggesting that this term was recovered not only with high precision, but also with minimal bias.

We repeated the same exercise, this time using parameters from the short version of Experiment 2 and simulating 72 trials per participant. Due to the lower number of simulated trials, the correlations here were lower. Still,  $\bar{\alpha} - \alpha$  was recoverable well above chance,  $r = .62, 95\% \text{ CI } [.54, .69], t(232) = 12.08, p < .001$ . Again, the model intercept was not significantly different from 0,  $t(232) = 1.64, p = .102$ , suggesting that the relationship between  $\alpha$  and  $\bar{\alpha}$  can be recovered without introducing a bias.

## Parameter Values

**Figure A3**

Parameter Value Distributions Across the Three Experiments



*Note.* True and believed effects of occlusion on visibility were strongly correlated, but participants' beliefs tended to underestimate the true effect of occlusion on visibility, with values closer to 1. Exp. = experiment. See the online article for the color version of this figure.

## Within-Condition Variability

In our experiments, incorrect responses were generally slower than correct responses. This was true both for “target present” responses (difference in seconds between false alarm and hit trials; Experiment 1:  $M = 0.33, 95\% \text{ CI } [0.26, 0.39]$ ; Experiment 2:  $M = 0.31, 95\% \text{ CI } [0.23, 0.38]$ ; Experiment 3:  $M = 0.39, 95\% \text{ CI } [0.32, 0.47]$ ), and for “target absent” responses (difference in seconds between miss and correct rejection trials; Experiment 1:  $M = 0.31, 95\% \text{ CI } [0.24, 0.37]$ ; Experiment 2:  $M = 0.37, 95\% \text{ CI } [0.31, 0.43]$ ; Experiment 3:  $M = 0.27, 95\% \text{ CI } [0.20, 0.35]$ ).

As presented in the main text, the model does not consistently account for these effects, sometimes even predicting that correct responses should be slower rather than faster than incorrect responses (“target present” responses in Experiment 1:  $M = 0.03, 95\% \text{ CI } [0.01, 0.05]$ ; Experiment 2:  $M = -0.05, 95\% \text{ CI } [-0.08, -0.03]$ ; Experiment 3:  $M = 0.07, 95\% \text{ CI } [0.05, 0.08]$ ; “target absent” responses in Experiment 1:  $M = -0.01, 95\% \text{ CI } [-0.03, 0.01]$ ; Experiment 2:  $M = 0.06, 95\% \text{ CI } [0.04, 0.08]$ ; Experiment 3:  $M = -0.07, 95\% \text{ CI } [-0.09, -0.04]$ ).

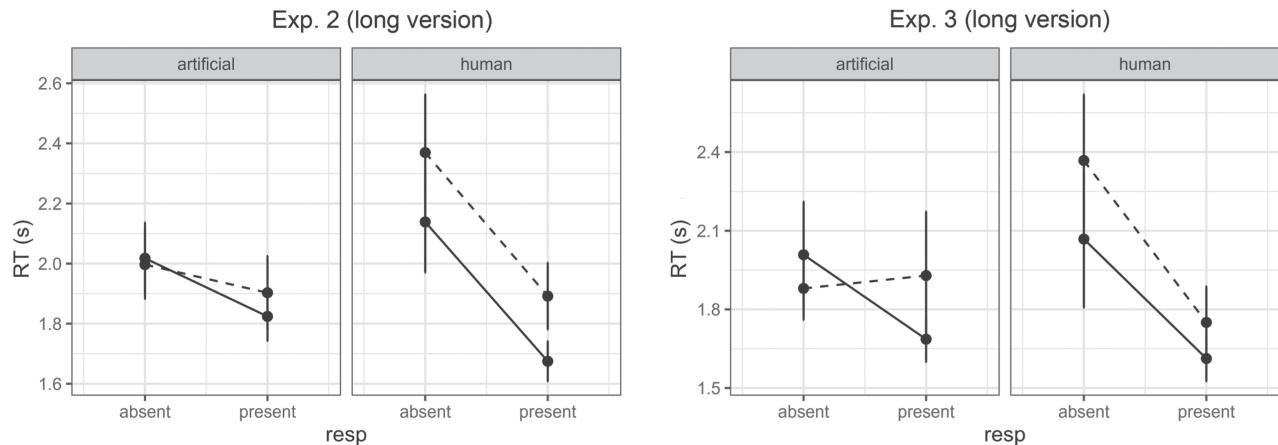
Of note, the presented model assumes that all trials within a condition are of the same difficulty (i.e., the visibility of the stimulus and the believed visibility of the stimulus are unchanged across trials). In the context of drift diffusion modeling, assuming that evidence accumulation varies between trials can account for slower error trials (Calder-Travis et al., 2024; Ratcliff & McKoon, 2008). Therefore, to incorporate intertrial variability into our model, we extended the model by assigning a random “difficulty” value  $x$  to each trial, sampled uniformly from  $x \in [-2, -1, 0, 1, 2]$ . Target visibility on a given trial was then defined as  $\alpha \eta^x \bar{\theta}$ , and beliefs about visibility on a given trial are  $\bar{\alpha} \bar{\eta}^x \bar{\theta}$ , with both  $\eta$  and  $\bar{\eta}$  in the range of  $[0, 1]$ . This way, lower values of  $\theta$  give rise to more pronounced variability in the true visibility of stimuli (with  $\theta = 1$  corresponding to no variability at all), and lower values of  $\bar{\theta}$  produce higher variability in beliefs about the visibility of stimuli. In this specification of the model, visibility and believed visibility are perfectly correlated across trials, but a third parameter can be introduced to control the alignment between the two.

We fitted the extended model to participants’ behavior in the long version of Experiments 2 and 3 (given the higher number of free parameters, using the long version ensured we had a reasonable number of data points per free parameter). As can be seen in Figure A4, the extended model successfully accounts for slower error trials both in “target present” and in “target absent” responses. Furthermore, inspection of the fitted model parameters suggests that participants underestimated the true variability in stimulus visibility across trials in Experiment 2 ( $\eta - \bar{\eta}, t(9) = -3.39, p = .008$ ), but not in Experiment 3, where the expected visibility of stimuli could be directly perceived in the reference stimuli ( $t(9) = -0.31, p = .764$ ).

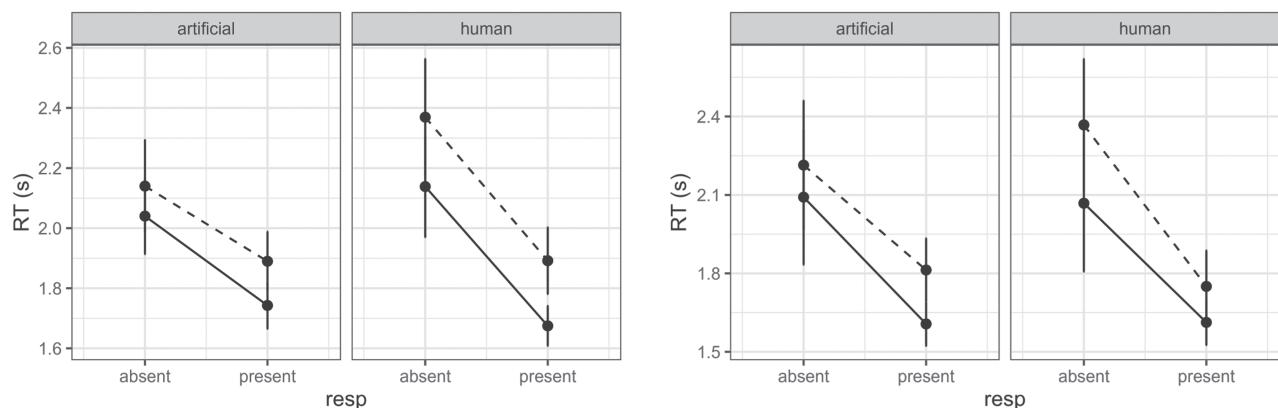
(Appendix continues)

**Figure A4***Association Between Decision and Decision Accuracy in Human and Artificial Data*

(A) basic model



(B) trial variability



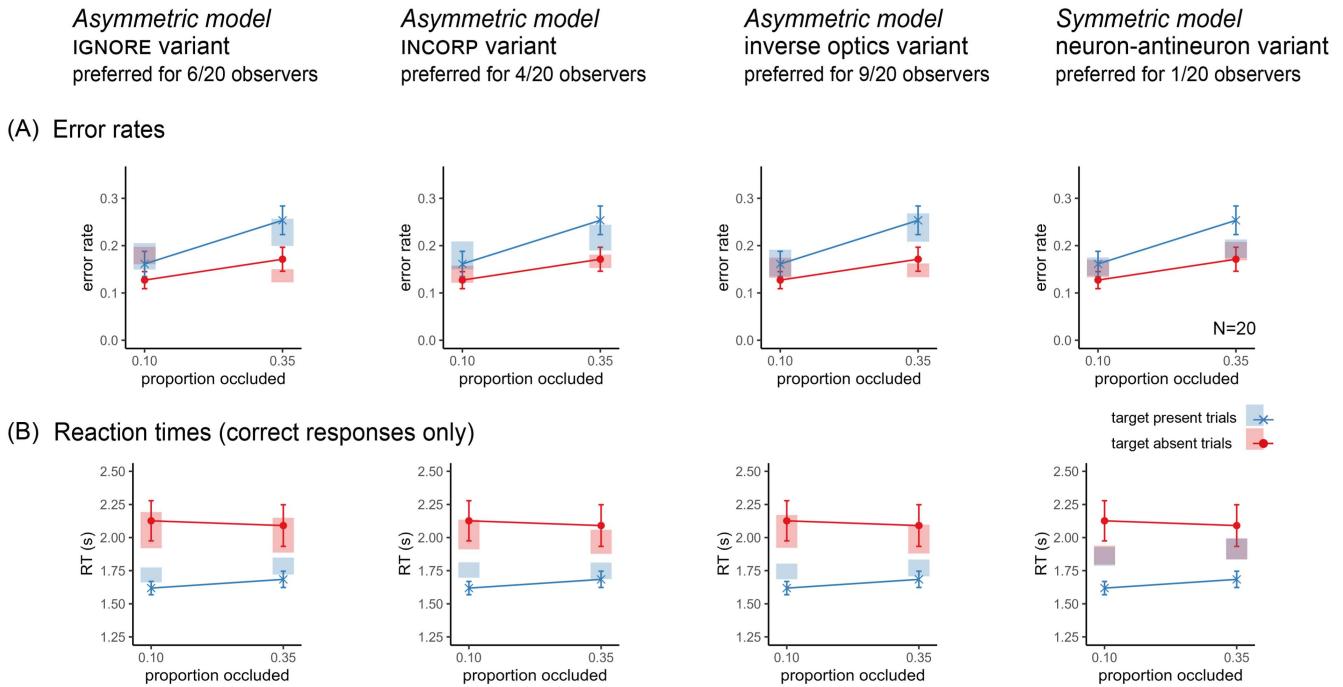
*Note.* Solid lines represent correct responses, dashed lines are incorrect responses. Panel A: the standard model, as presented in the main text. Panel B: the extended model, with intertrial variability. Exp. = experiment; RT = reaction time; resp = response.

(Appendix continues)

## Model Fits for All Model Variants

**Figure A5**

*Empirical Data and Model Predictions for the Four Model Variants, Focusing on Data From the Two Long Experiments*



*Note.* Same conventions as Figure 6. INCORP = incorporate; RT = reaction time. See the online article for the color version of this figure.

## Reverse Correlation of Sensor Activation Sequences

We used the fitted parameters from the full asymmetric and symmetric models to simulate 64 trials from the participants who took part in Experiment 2, producing two artificial data sets. Critically, we recorded not only trial-wise measures such as decision and response time but also the exact sequence of sensor activation per trial. We then subjected these sequences to a reverse correlation analysis, applying the same preprocessing and analysis steps as we did for human data.

Notably, this is reverse correlation over latent internal variables, rather than over fluctuations in stimulus visibility as is the case for human observers where internal states had to be inferred rather than directly observed. While it is reasonable to assume that the

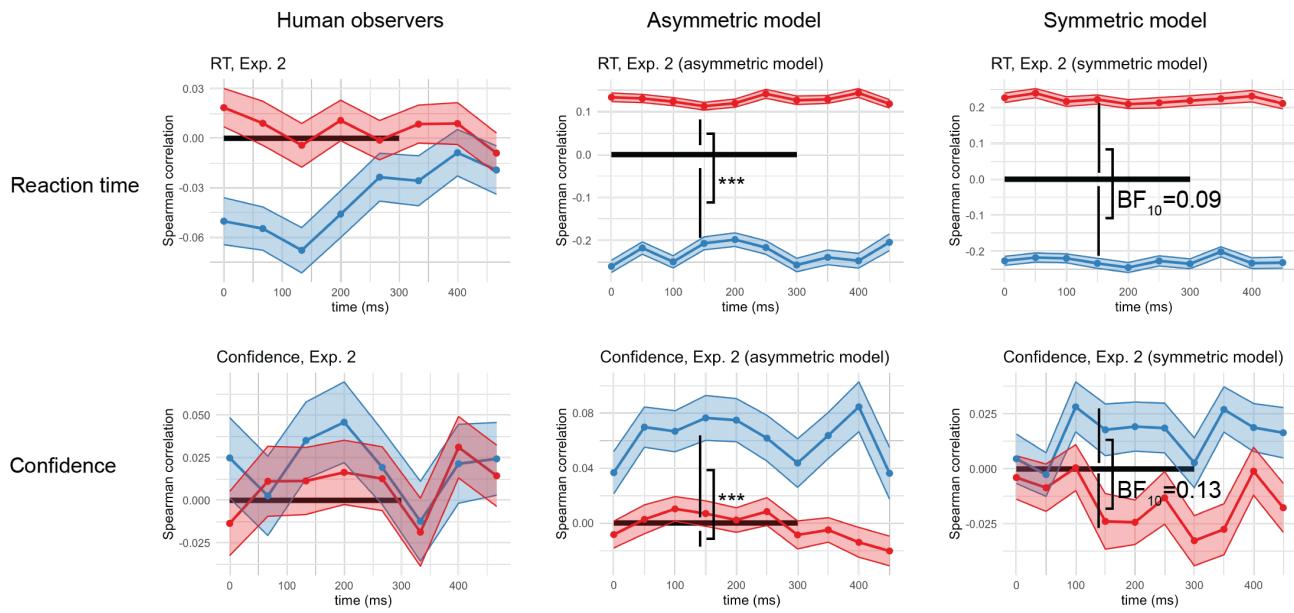
probability of sensor activation scales with the visibility of the external stimulus, the exact function linking the two can take many forms, which can give rise to different predictions. To keep our analysis neutral with respect to such assumptions, we opted to run the analysis on the activations themselves.

Reverse correlation kernels from the asymmetric model showed the same qualitative asymmetry found in human data, whereby perceptual evidence contributed more to decision time and decision confidence when a target was present. This asymmetry was not observed, however, when applied to simulated data from the symmetric model (using the difference between presence and absence sensor activations as a measure of perceptual evidence).

(Appendix continues)

**Figure A6**

*Reverse Correlation Kernels From Human Data (Computed Over Normalized Target–Stimulus Similarity Measures) and From Simulated Data (Computed Over Sequences of Sensor Activations)*



*Note.* Only the asymmetric model produces an asymmetry between the effect of perceptual evidence on decisions, when a target is present versus absent. Same conventions as Figure 3D and 3E. Significance stars and Bayes factors refer to a comparison between target-present and target-absent effects in the first 300 ms of the trial. RT = reaction time; Exp. = experiment; BF = Bayes factor. See the online article for the color version of this figure.

\*\*\*  $p < .001$ .

Received July 26, 2024  
 Revision received December 9, 2024  
 Accepted January 26, 2025 ■