The role of counterfactual visibility in inference about absence

Matan Mazor[1,2], **Rani Moran**[3,4] **& Clare Press**[1,5]
[1] Birkbeck, University of London, [2] University of Oxford, [3] Queen Mary, University of London,
[4] Max Planck UCL Centre for Computational Psychiatry and Aging Research, [5] University College London

Abstract

We provide a generalized, normative model of visual detection that accounts for key asymmetries between decisions about presence and about absence. In our model, decisions about presence are made based on the visibility of presented stimuli, but decisions about absence are made based on counterfactual visibility: beliefs about the degree to which a stimulus would have been visible if present. Behavioral patterns in visual detection experiments under different levels of partial occlusion validate key model predictions. Specifically, we find that unlike decisions about presence, the confidence and speed of decisions about absence are largely independent of perceptual evidence, but are sensitive to the counterfactual visibility of absent stimuli. Finally, we reveal robust individual differences in counterfactual perception, with some participants systematically incorporating counterfactual visibility into perceptual decisions in a different fashion from others. We discuss implications for the varied and inferential nature of visual perception more broadly.

Keywords: perceptual decisions; counterfactual reasoning; absence; metacognition; Bayesian modeling; ideal observer

Introduction

After checking Taylor Swift's Wikipedia page, we are confident that she hasn't announced her retirement from music. If she had, it would have been mentioned on her page. We also checked Russian cellist Natalia Gutman's page and didn't see any mention of a similar announcement, but we are not so sure she hasn't 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.

More generally, inferences about the negation of a hypothesis (H) depend on our belief in the probability that we would observe evidence (E) if H were true ($p(E|H)$) (Oaksford & Hahn, 2004; Walton, 1992, 2010). In other words, we believe that something is not true (for example, that Taylor Swift hasn't announced her retirement from music) when we believe that "if it were true, we would have heard about it by now" (Goldberg, 2011).

Here we ask whether a similar principle is at play within perception. According to inferential, "inverse optics" accounts, perception is the inference of latent external causes based on noisy sensory data (Alhazen & Smith, 2001; Friston, 2010; Gershman, Vul, & Tenenbaum, 2012; Helmholtz, 1866). For such inference to be rational, evidence, or its absence, is interpreted in light of its relative likelihood under competing world states. As a result, perceptual decisions in the absence of a stimulus, their biases, timing, and confidence, should be affected not only by visibility, but also by *counterfactual visibility* – implicit beliefs about the visibility of hypothetical stimuli that are not in fact present. Here we ask whether counterfactual visibility is in fact used when making perceptual decisions.

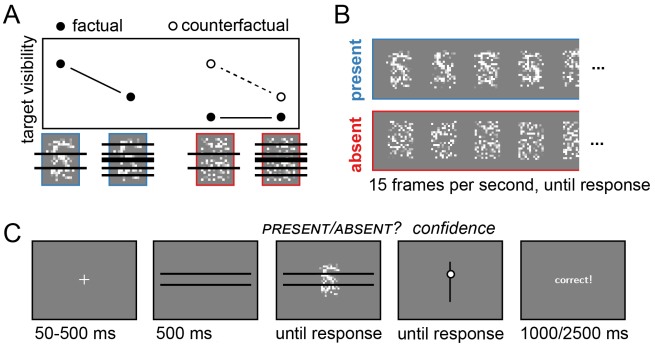

Figure 1: Rationale and experimental design. A) occluding more of a target stimulus (for example, the letter S) decreases its visibility (black markers). Occlusion has no effect on target visibility when the target is absent, but it affects counterfactual visibility (white markers): the expected visibility of the target, had it been present. B) example frames from target-present (blue) and target-absent (red) trials. C) trial structure. Participants performed two 32-trial detection 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 randomised between participants. On different trials we occluded a random subset of 2 or 6 pixel rows.

Results

Overall, we ran a series of three pre-registered and two exploratory experiments. Our key results were consistent across all experiments. Due to space limitations, we describe here the results from the one experiment in the series in which we collected subjective confidence ratings. Experiment demos, reproducible analysis code, links to pre-registrations and the full pre-registered analysis from all experiments are available at github.com/matanmazor/counterfactualVisibility.

Perceptual detection under partial occlusion

252 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 (Fig. 1B). On different trials, either 2 or 6 rows of pixels were occluded by randomly positioned lines (Fig. 1A). 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". Importantly, the stimulus remained on the screen, refreshing at 15 frames per second, until a response was made. After making a decision, participants rated their confidence on an analog scale (Fig. 1C). Based on our pre-registered exclusion criteria (below-chance accuracy or more than 25% of reaction times below 100 ms or above 5000 ms), we excluded 18 participants, leaving 234 for the main analysis.

Presence-absence asymmetries After exclusion, mean accuracy in the main experiment was 0.81 (SD=0.03). Participants were biased to report target absence (0.57 of all responses, SD= 0.08). Consistent with the response time and confidence profile of detection tasks (Mazor, 2021; Mazor & Fleming, 2022; Mazor, Maimon-Mor, Charles, & Fleming, 2023; Mazor, Moran, & Fleming, 2021), response times were significantly shorter in "target present" compared to "target absent" responses (1.98 vs 2.40 seconds; $t(233) = -15.18$, $p < .001$; Fig. 2A), and confidence was higher in decisions about presence compared to absence (0.93 vs. 0.92 on a 0.5-1 (guess to full certainty) scale; $t(233) = 2.92$, $p = .004$; Fig. 2B).

Exploratory 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 how stimulus-target similarity or dissimilarity (quantified as the Pearson correlation between unoccluded pixels and their corresponding pixels in the target letter, statistically controlling for the proportion of pure-noise and hidden pixels in the frame) contributed to reaction times in decisions about presence and absence.

Following previous reverse correlation studies of decision confidence (Mazor et al., 2023; Zylberberg, Barttfeld, & Sigman, 2012), we focused our analysis on the first 300 ms of stimulus presentation, and extracted, per frame and per trial, the mean similarity between the visual noise and the target letter in these first frames. We then computed the Spearman correlation between these trial-wise similarity measures and reaction times and confidence, focusing our analysis on correct responses only (see Fig. 2).

Higher levels of stimulus-target similarity in the first 300 ms of the trial made participants quicker to detect the target letter when it was present ($t(250) = -7.78$, $p < .001$; Fig. 2C), and made them more confident in their correct "target present" decisions ($t(234) = 3.15$, $p = .002$; Fig. 2D). In contrast, stimulus-target similarity had no effect on

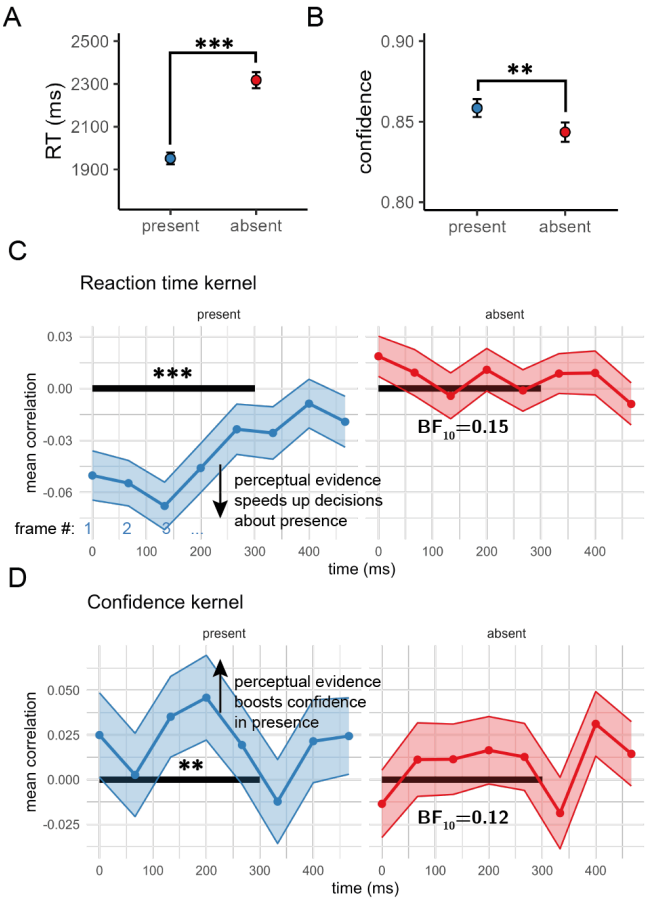

Figure 2: Behavioural results. Detection asymmetries in reaction times (A) and subjective confidence (B). Reverse correlation reveals detection asymmetries in the weighting of perceptual evidence in reaction time (C) and confidence (D). Individual points represent the mean correlation value for individual frames. Error bars and shaded areas represent the standard error of the mean. Black lines represent the first 300 ms of the trial. **: p<0.01,***:p<0.001.

the speed ($t(250) = 1.25$, $p = .214$; $BF_{10} = 0.15$) or confidence ($t(242) = 0.79$, $p = .432$; $BF_{10} = 0.12$) of correct decisions about absence. Moreover, the effect of stimulus-target similarity was stronger in "target present" compared to "target absent" responses for both reaction time ($t(249) = 4.62$, $p < .001$) and confidence ($t(232) = 3.13$, $p = .002$). Unlike decisions about presence, which were driven by the similarity of the stimulus to the target letter, the speed and confidence of decisions about absence was not based on dissimilarity to the target letter.

An ideal observer model of perceptual detection

Presence-absence asymmetries in reaction time 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. This explanation is consis-

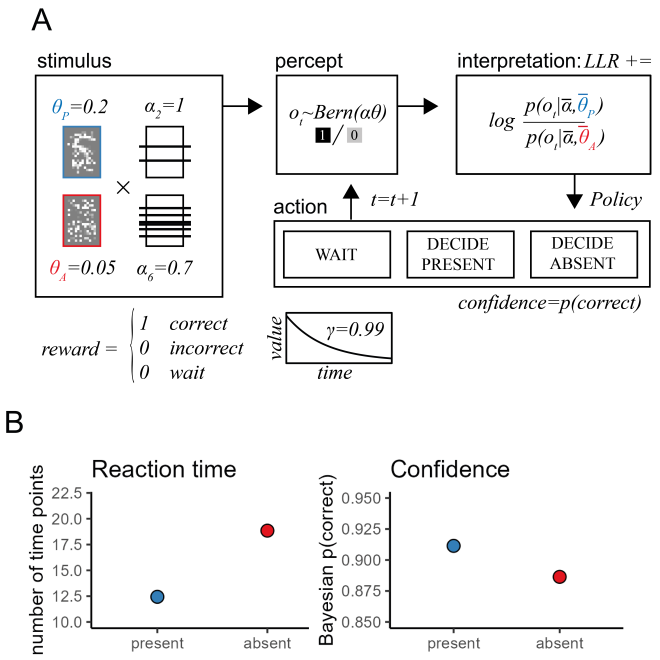

Figure 3: An ideal observer model of detection. A: model specification. B: predictions for an ideal observer.

tent with our reverse correlation results, where perceptual evidence drove confidence and decision times in decisions about presence, but not absence. To formulate this asymmetry in the availability of evidence, we present a Partially Observed Markov Decision Process (POMDP, Littman, 2009) model of perceptual detection (Fig. 3A). We begin by presenting the model in its simplest form, before introducing the additional effects of occlusion. Crucially, our motivation here is to ask about the evidence structure that renders participants' behaviour rational (Anderson, 1990). As we show, asymmetries in decision time and decision confidence are borne out of rational evidence accumulation when the value of positive and negative evidence is itself asymmetrical.

A POMDP is a 7-tuple: $<\mathcal{S},\mathcal{A},\mathcal{T},\Omega,O,\nabla,\gamma>$. The state space $\mathcal{S}$ comprises two states describing target presence or absence and two additional states for trial endings: correct and incorrect. The action space $\mathcal{A}$ has three possible actions: "wait", "decide present", and "decide absent". The transition function $\mathcal{T}:(\mathcal{S},\mathcal{A})\rightarrow\mathcal{S}$ specifies the effect of actions on state transitions. "wait" maps states to themselves, and deciding maps states to the "correct" or "incorrect" states depending on the accuracy of the decision. Ω is the set of possible observations. We assume these are $[0,1]$, that is, perceptual evidence has a binary form. $O:\mathcal{S}\rightarrow P(\Omega)$ is a probabilistic function from states to observations, which we describe in more detail below. $\nabla:\mathcal{S}\rightarrow R$ 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 γ affects the subjective value of anticipated rewards. We set $\gamma:=0.99$, meaning that a reward obtained in the next time point is worth 0.99 of its worth if obtained now.

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

$$\theta := \begin{cases} 0.05 & \text{absent} \\ 0.2 & \text{present} \end{cases}$$

Importantly, our qualitative results of slower and less confident decisions about absence hold for any choice of θ such that $0<\theta_{absent}<\theta_{present}<0.5$, because, for values within this range, positive evidence (that is, sampling a 1) is more informative than negative evidence (that is, 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 log likelihood ratio *LLR* between target presence and absence, which they update following each sample.

$$LLR_t = \sum_{i=1}^{t} log \frac{p(o_i|\bar{\theta}_{presence})}{p(o_i|\bar{\theta}_{absence})}$$

Where

$$p(o_i = 1|\bar{\theta}) = \bar{\theta}$$

With $\bar{\theta}$ being the assumed value of θ 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:

$$p(present|O_t) = \frac{e^{LLR_t}}{1+e^{LLR_t}}$$

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 $p(present|O)>0.5$, the probability of being correct at that time point is:

$$p(correct|DECIDE,O_t) = max(p(present|O_t),1-p(present|O_t))$$

When following the optimal policy, the expected value at time point t equals the probability of being correct, unless the value of additional evidence outweighs the discount factor γ:

$$E(V|O_t) = max(p(correct|O_t),$$
$$p(1|O_t)\gamma E(V|[O_t,1]) + p(0|O_t)\gamma E(V|[O_t,0]))$$

Where $E(V|[O_t,1])$ is the expected value at time point $t+1$, assuming the next sample is 1, and $p(1|O_t)$ is the probability that the next sample will be 1, marginalized over target presence and absence (similar for 0).

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

To find the optimal policy (the one that maximizes the Bellman equation), we used backward induction with a horizon of 100 time points (Callaway, Griffiths, Norman, & Zhang, 2023; Puterman, 2014). We then simulated 4000 trials to obtain predictions for a rational decision-maker.

An ideal observer model successfully reproduced the behavioural asymmetries in decision times and confidence: response times were shorter for correct "target present" decisions (mean = 12.81 time points until decision) than for correct "target absent" decisions (mean = 18.25 time points). Second, subjective confidence was higher in correct decisions about presence (mean = 0.91) than absence (mean = 0.89; Fig. 3B). Finally, and in line with participants' behaviour, the model was biased to report target absence (0.52 of all simulated responses) despite having an accurate prior about a letter being present in exactly half of the trials. These behavioural asymmetries emerged not due to an asymmetric prior or incentive structure, but due to asymmetries in the likelihood function going from world states to perceptual input.

Modeling 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 $\theta_{present}$ and θ_{absent}, α is paralleled by a metacognitive variable, $\bar{\alpha}$, which corresponds to participants' beliefs about the 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 (α) 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 log likelihood ratio between target presence and absence following a 1, is unaffected by beliefs about the effects of occlusion on visibility, $\bar{\alpha}$:

$$LLR_{[1]} = log\frac{p(1|present)}{p(1|absent)} = log\frac{\bar{\alpha}\bar{\theta}_{present}}{\bar{\alpha}\bar{\theta}_{absent}} = log\frac{\bar{\theta}_{present}}{\bar{\theta}_{absent}}$$

And third, the informativeness of obtaining negative evidence, quantified as the log likelihood ratio 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:

$$|LLR_{[0]}| = |log\frac{p(0|present)}{p(0|absent)}| =$$
$$|log\frac{1-\bar{\alpha}\bar{\theta}_{present}}{1-\bar{\alpha}\bar{\theta}_{absent}}| < |log\frac{1-\bar{\theta}_{present}}{1-\bar{\theta}_{absent}}|$$

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 following simulation we had two occlusion levels; one where $\alpha = 1$ (easy condition) and one where $\alpha = 0.7$ (hard condition). To illustrate the effects of beliefs about visibility on perceptual decisions, we present the results of two simulated agents: V_{incorp} is an ideal observer who incorporates accurate beliefs about the expected effect of occlusion on visibility into perceptual decisions ($\bar{\alpha} = \alpha$), and V_{ignore} is an observer who ignores the expected effects of occlusion on visibility, interpreting perceptual evidence similarly in both levels of occlusion ($\bar{\alpha} = 0.85$ for both hard and easy conditions). For both agents, we found the optimal policy (given their beliefs) using backward induction, and simulated 4000 trials to obtain predictions (Fig. 4, left and middle columns).

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 Fig. 4). While occluding more of the display made V_{incorp} commit more false-alarms, it made V_{ignore} make fewer of them. V_{incorp}'s decisions about absence were slower when more of the display was occluded, whereas V_{ignore}'s decisions about absence were faster. Finally, V_{incorp} was less confident in decisions about absence when more of the display was occluded, but this was not true of V_{ignore}. Together, both the size and direction of occlusion effects on perceptual decisions were dependent on meta-perceptual knowledge about the influence of occlusion on visibility, or the incorporation of such knowledge into perceptual decisions.

Measuring Occlusion effects Equipped with an ideal-observer model of perceptual detection under partial stimulus occlusion, we now return to describe human data. As expected, hit rate (the probability of a "target present" response in target-present trials) was reduced by occlusion ($t(233) = 11.83$, $p < .001$), with a mean hit rate of 0.78 (SD= 0.11) when 2 rows of pixels were occluded, versus 0.66 (SD= 0.12) when 6 rows of pixels were occluded. Unsurprisingly, occluding more of the target made it more difficult to spot.

Next, we asked whether stimulus occlusion affected the timing and confidence with which participants detected letters. Correct "target present" decisions were slowed down by pixel occlusion (1.92 vs 2.05 seconds for 2 or 6 occluded rows; $t(233) = -5.13$, $p < .001$). Similarly, confidence in correct "target present" decisions was lower when more of the display was occluded (0.94 vs. 0.92 on a 0.5-1 scale; $t(233) = 9.87$, $p < .001$).

Having established that occlusion affected stimulus visibility, making responses slower, less accurate, and less certain, when a target was present, we examined the effects of

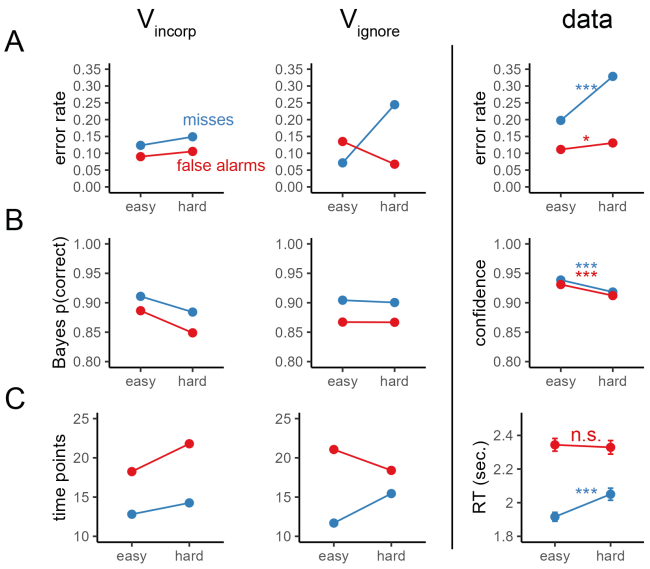

Figure 4: Simulated predictions and empirical data for the effects of occlusion on error rate (panel A), subjective confidence (panel B), and reaction time (panel C). Agent V-incorp used metacognitive knowledge of the effect of occlusion on visibility, but agent V-ignore did not. Error bars represent the standard error of the mean. *: p<0.05 ***: p<0.001

occlusion on detection responses in the absence of a target. As per our modelling above, the effects of occlusion on decisions about absence reveal not only how occlusion affects visibility (modelled as α), but also how it affects counterfactual visibility: the expected visibility of a stimulus that is in fact absent (closely related to $\bar{\alpha}$ in the model). Consistent with the predictions for V_{incorp}, but not V_{ignore}, occluding more of the display resulted in an increase in the false-alarm rate (0.13 versus 0.15 for 2 or 6 occluded rows, respectively; $t(233) = -2.26$, $p = .025$): participants were more likely to accept that they might have missed the target when more of the stimulus was occluded. Similarly, occlusion had a negative effect on confidence in absence (0.93 vs. 0.91; $t(233) = 10.54$, $p < .001$). However, and in contrast to the predictions for both model variants V_{incorp} and V_{ignore}, occlusion had no effect, positive or negative, on the speed of decisions about absence (2.34 vs 2.33 seconds for 2 or 6 occluded rows; $t(233) = 0.79$, $p = .429$. $BF_{10} = 9.97 \times 10^{-2}$). Moreover, the effect of occlusion on response times was significantly stronger in "target present" compared to "target absent" responses ($t(233) = -4.68$, $p < .001$), also when incorporating incorrect responses into the analysis ($t(233) = -3.89$, $p < .001$).

Individual differences in counterfactual perception 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 model variants: variant V_{incorp} predicted a positive effect, and variant V_{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_{incorp}, incorporating counterfactual 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_{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 makes two unique predictions. First, despite a group-level null effect, some individual participants should show reliable effects of occlusion on "target absent" reaction times: negative for some participants, and positive for others. And second, participants who slow down when more of the display is occluded should, paradoxically, make more false alarms in this condition (see Fig. 4, red lines in panels A and C). Notably, this correlation is the exact opposite of what is expected from a speed-accuracy trade-off.

To test the first hypothesis, we collected a large number of test trials (between 672 and 894) from a randomly-chosen subset of participants who took part in previous experiments. We present here the combined results from two cohorts of participants. The first cohort participated in a long version of the experiment described above, without confidence ratings. For the second cohort, the central stimulus was flanked by two target-present stimuli, to be used as a visual reference for the effect of occlusion on visibility.

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 1 out 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 8, split exactly half-half between significant positive effects (more consistent with model variant V_{incorp}) and significant negative effects (more consistent with model variant V_{ignore}): much higher than the 1/20 probability expected by chance alone ($p < 0.001$ in a binomial test against $p = 0.05$).

As a more sensitive test of effect reliability, we employed the non-parametric *sign-consistency test* (Yaron, Faivre, Mudrik, & Mazor, 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 Fig. 5). 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-

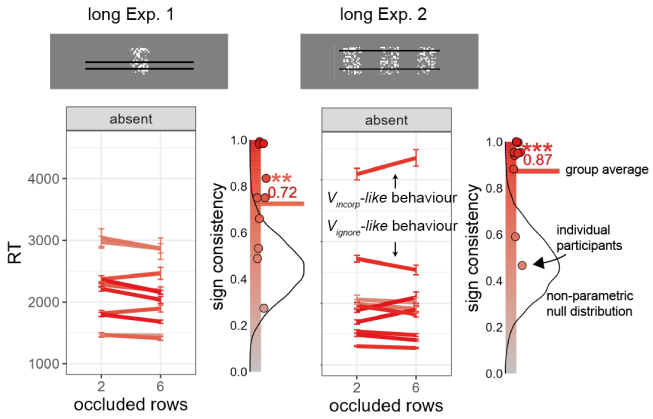

Figure 5: Sign consistency results in Exp. 1 (left) and 2 (right). Within each panel, we present median RT as a function of occlusion level for each participant on the left. Color saturation indicates sign-consistency. On the right, we present individual sign-consistency scores as circles, alongside the group-average sign consistency score (horizontal line), overlaid on top of the non-parametric 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. **: p<0.01, ***: p<0.001

chance sign-consistency in the effects of occlusion on target-absent reaction times (Exp. 1: sign consistency=0.72, p= .003; Exp. 2: sign consistency=0.86, $p<$ .001; see Fig. 5). Moreover, target-absent sign-consistency scores were not significantly different from, and numerically higher than, target-present sign-consistency scores (Exp. 1: sign consistency=0.63; Exp. 2: sign consistency=0.76). An effect of counterfactual visibility on "target absent" response times was not absent: it was masked by differences between individual participants who systematically exhibit opposing influences.

Turning to the second prediction of the population-mixture model, we measured the correlation between the effect of occlusion on "target absent" decision times and the false-alarm rate. In both cohorts, we find that subjects who waited for longer before inferring absence when more of the display was occluded committed more false alarms in the high-occlusion condition (Exp. 1: r =.41; Exp. 2: r =.69; analyzed together: $r = .55$, 95% CI $[.14, .80]$, $t(18) = 2.80$, $p = .012$). Importantly, this is not due to variability in the perceptual effect of occlusion on visibility (for comparison, we find no reliable correlation between the effect of occlusion on target-present reaction times and the hit rate; $r = -.20$, 95% CI $[-.59, .26]$, $t(18) = -0.89$, $p = .388$). Instead, we argue that this correlation reflects variability in the use of counterfactual visibility to inform perceptual decisions in the absence of a target.

Discussion

Occlusion affects not only the visibility of objects, but also the counterfactual visibility of absent objects: how visible they would have been had they been present. Here, to pinpoint the roles of counterfactual visibility in perceptual decision making, we asked whether occlusion had similar effects on perceiving stimuli and their absence. Below we summarise the three main contributions of this paper.

First, we provide an ideal observer model of perceptual detection. This model traces presence-absence asymmetries in reaction time and confidence to asymmetries in the information value of positive and negative evidence in a detection setting. The model also formalises the argument that metacognitive beliefs about perception have a key role in decisions about absence (Kanai, Walsh, & Tseng, 2010; Mazor, 2021).

Second, we find evidence for the incorporation of counterfactual visibility into perceptual decisions in the absence of a target, as well as subjective confidence in such decisions. Using reverse correlation, we further show that timing and confidence have strikingly different origins in decisions about presence and absence: available perceptual evidence versus beliefs about the availability of counterfactual evidence.

Finally, the effect of occlusion on "target absent" decision times was reliably variable across individuals. Computational modelling suggests that this variability is related to individual differences in beliefs about perception, or in the incorporation of such beliefs into perceptual decisions, raising questions regarding associations with reasoning about counterfactuals outside perception (Byrne & Tasso, 1999; Hsu, Horng, Griffiths, & Chater, 2017), susceptibility to expectation effects on perception (Kok, Brouwer, Gerven, & Lange, 2013; Powers, Mathys, & Corlett, 2017; Press, Kok, & Yon, 2020), and meta-perceptual knowledge (Levin & Angelone, 2008; Recht, Gardelle, & Mamassian, 2021; Scholl, Simons, & Levin, 2004).

Our results fit within the broader project of understanding perception as probabilistic inference on noisy sensory data. Much focus has been placed on the role of prior expectations in perceptual inference (Kok et al., 2013; Press et al., 2020; Summerfield & Egner, 2009; Yon, Zainzinger, Lange, Eimer, & Press, 2021), with important discussions regarding the (im)penetrability of visual perception to such effects from cognition (Firestone & Scholl, 2016; Pylyshyn, 1999). Here we focus on the other component of Bayesian reasoning, often neglected in such discussions: the likelihood function going from world states to sensory input. Unlike prior expectations about the world (e.g., the probability that a letter will be present), these likelihood functions describe the perceptual system itself (e.g., the probability that I would perceive the letter when it is present). As we show here, such beliefs affect not only metacognitive confidence ratings, but also decision times and decision criteria of the detection judgments themselves, revealing a complex web of interactions between perception, cognition, and metacognition.

Acknowledgements

This work was supported by a European Research Council (ERC) consolidator grant (101001592) under the European

Union's Horizon 2020 research and innovation programme, awarded to CP. MM is supported by a post doctoral research fellowship from All Souls College at the University of Oxford. We thank Daniel Yon and Mathias Sablé-Meyer for useful feedback on previous versions of this paper.

References

Alhazen, & 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.

Anderson, J. R. (1990). *The adaptive character of thought*. New York: Psychology Press. http://doi.org/10.4324/9780203771730

Byrne, R. M. J., & Tasso, A. (1999). Deductive reasoning with factual, possible, and counterfactual conditionals. *Memory & Cognition*, *27*(4), 726–740. http://doi.org/10.3758/BF03211565

Callaway, F., Griffiths, T. L., Norman, K. A., & Zhang, Q. (2023). Optimal metacognitive control of memory recall. *Psychological Review*, No Pagination Specified–No Pagination Specified. http://doi.org/10.1037/rev0000441

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, *39*, e229. http://doi.org/10.1017/S0140525X15000965

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. http://doi.org/10.1038/nrn2787

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, *24*(1), 1–24. http://doi.org/10.1162/NECO_a_00226

Goldberg, S. (2011). If that were true i would have heard about it by now. *Social Epistemology: Essential Readings*, 92–108.

Helmholtz, H. von. (1866). Concerning the perceptions in general. *Treatise on Physiological Optics,*.

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*(S5), 1155–1167. http://doi.org/10.1111/cogs.12356

Kanai, R., Walsh, V., & Tseng, C. (2010). Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness. *Consciousness and Cognition*, *19*(4), 1045–1057. http://doi.org/10.1016/j.concog.2010.06.003

Kok, P., Brouwer, G. J., Gerven, M. A. J. van, & Lange, F. P. de. (2013). Prior Expectations Bias Sensory Representations in Visual Cortex. *Journal of Neuroscience*, *33*(41), 16275–16284. http://doi.org/10.1523/JNEUROSCI.0742-13.2013

Levin, D. T., & Angelone, B. L. (2008). The visual metacognition questionnaire: A measure of intuitions about vision. *The American Journal of Psychology*, *121*(3), 451–472. http://doi.org/10.2307/20445476

Littman, M. L. (2009). A tutorial on partially observable markov decision processes. *Journal of Mathematical Psychology*, *53*(3), 119–125. http://doi.org/10.1016/j.jmp.2009.01.005

Mazor, M. (2021). Inference about absence as a window into the mental self-model.

Mazor, M., & Fleming, S. M. (2022). Efficient search termination without task experience. *Journal of Experimental Psychology: General*.

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*, 1–30.

Mazor, M., Moran, R., & Fleming, S. M. (2021). Metacognitive asymmetries in visual perception. *Neuroscience of Consciousness*, *2021*(1). http://doi.org/10.1093/nc/niab025

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.

Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, *357*(6351), 596–600. http://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. http://doi.org/10.1016/j.tics.2019.11.003

Puterman, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Pylyshyn, Z. (1999). Is vision continuous with cognition?: The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, *22*(3), 341–365. http://doi.org/10.1017/S0140525X99002022

Recht, S., Gardelle, V. de, & Mamassian, P. (2021). Metacognitive blindness in temporal selection during the deployment of spatial attention. *Cognition*, *216*, 104864. http://doi.org/10.1016/j.cognition.2021.104864

Scholl, B. J., Simons, D. J., & Levin, D. T. (2004). *" change blindness" blindness: An implicit measure of a metacognitive error*. na.

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, *13*(9), 403–409. http://doi.org/10.1016/j.tics.2009.06.003

Walton, D. (1992). Nonfallacious arguments from ignorance. *American Philosophical Quarterly*, *29*(4), 381–387.

Walton, D. (2010). *Arguments from ignorance*. Penn State Press.

Yaron, I., Faivre, N., Mudrik, L., & Mazor, M. (2023). Individual differences do not mask effects of unconscious processing.

Yon, D., Zainzinger, V., Lange, F. P. de, Eimer, M., & Press,

C. (2021). Action biases perceptual decisions toward expected outcomes. *Journal of Experimental Psychology: General*, *150*(6), 1225.

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*. Retrieved from https://www.frontiersin.org/articles/10.3389/fnint.2012.00079